NOTE

This Microfiche meets the quality requirements of the American National Standard for Microfiche of Documents (ANSI PH5.9-1975). However, because of imperfections in the document from which it was made, it contains images that are illegible or difficult to read. The source document for this microfiche is the best copy that could be obtained prior to filming.

CIS

Congressional Information Service, Inc.
Washington, D.C.

Congress of the United States
Office of Technology Assessment

MAPPING OUR GENES
Federal Genome Projects: How Vast? How Fast?

Transcript of Workshop
"Costs of Human Genome Projects"

August 7, 1987

Discussions from this workshop were used in preparing OTA's final assessment report. OTA makes this transcript available for the use of readers desiring a more detailed discussion of an issue than can normally be accommodated in our final report. This transcript has not been reviewed or approved by the Technology Assessment Board. The views expressed at the workshop are those of the participants and do not necessarily reflect the views of OTA, the assessment advisory panel or the Technology Assessment Board.

J952-6

This document can be ordered in either paper or microfiche format from:

National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22161
Telephone: (703) 487-4650
Telex: 89-9405

CONGRESS OF THE UNITED STATES

- - -

OFFICE OF TECHNOLOGY ASSESSMENT

- - -

WORKSHOP: "COSTS OF HUMAN GENOME PROJECTS"

- - -

AUGUST 7, 1987

- - -

The Workshop was held in Conference Room D, 600 Pennsylvania Avenue, S.E., Washington, D.C., at 9:30 a.m., Paul Berg, Chairman, presiding.

- - -

PRESENT:

PAUL BERG, Stanford University, Stanford, California

CHRISTIAN BURKS, National Laboratories, Los Alamos, New Mexico

GEORGE F. CAHILL, Howard Hughes Medical Institute, Bethesda, MD

ANTHONY V. CARRANO, National Laboratory, Livermore, California

HELEN DONIS-KELLER, Collaborative Research, Inc., Bedford, Massachusetts

PRESENT: (continued)

WALTER GILBERT, Biological Laboratories, Harvard University, Cambridge, Massachusetts

LEROY HOOD, California Institute of Technology, Pasadena, California

KEITH McKENNEY, National Bureau of Standards, Gaithersburg, Maryland

MARK L. PEARSON, E.I. duPont de Nemours & Company, Wilmington, Delaware

ROBERT E. STEVENSON, American Type Culture Collection Rockville, Maryland

JOHN SULSTON, MRC Laboratory of Molecular Biology, Cambridge, England

JAMES D. WATSON, Cold Spring Harbor Laboratory, Long Island, New York

ALSO PRESENT FROM THE OTA STAFF:

ROBERT COOK-DEEGAN

DAVID H. GUSTON

GRETCHEN KOLRUD

JACQUELINE COURTEAU

PATRICIA HOBEN

ROGER HERDMAN

INDEX

Page

Opening Statement by the Chairman 4

Charge to Panel by Mr. Cook-Deegan 6

Discussion of RFLP Map 18

Discussion of Physical Map 74

Discussion of Technological Developments 148

Discussion of Organization 251

PROCEEDINGS

Time: 9:31 a.m.

CHAIRMAN BERG: I'm going to call this meeting to order a couple of minutes early. As I understand the agenda, we're going to have an opening -- several opening statements, the first of which from Ms. Kolrud.

MS. KOLRUD: Thank you. I'm Gretchen Kolrud. I'm the Manager of the Biological Applications Program here at OTA, and on behalf of OTA I'd like to welcome you all here today and to thank you in advance for your assistance in our assessment of the pros and cons of various approaches to mapping and/or sequencing the human genome.

I'd like to make a couple of introductory remarks about the Office as contacts for your work, since the Office of Technology Assessment is not exactly a household word. The Office is one of four Congressional support agencies, the others being the Library of Congress established in 1800, the General Accounting office established in 1923; the Congressional Budget Office established in 1974, and OTA itself was established in 1972 but did not become operational until 1974. So you're working with a relatively young agency.

At least two factors led to the establishment of the Office. One was the environmental concern of the 1960s typified in books like Silent Spring and other public expressions of concern over the indirect as well as the direct consequences of new technology. It was felt that these should be considered and assessed before decisions were made with respect to the establishment of these technologies.

Another concern that led to the establishment of the Office was concern on the part of the Congress that it had inadequate analytical support power relative to the Executive Branch. As you know, there's supposed to be a balance of powers between our branches of government, and you certainly can't expect executive agencies such as NASA to tell you why you ought not to build a space shuttle.

So the Congress felt that it needed analytic support power, and established the Office of Technology Assessment for that purpose. The Office is a little bit unusual in the respect with which both legislative factors and other factors have been established to ensure that all points of view are represented in its work.

It has a Technology Assessment Board composed of six Senators and six Representatives, equally divided by party, which is almost unique among Congressional committees. Traditional mechanisms have been established to help ensure that as well. One of these is the very extensive use of advisory panels and workshops such as these to ensure that many different points of view are brought to bear on a subject.

Since you have all been selected to represent a variety of viewpoints, we hope that today you will all speak up and represent those viewpoints so that you can assist us in this process.

With that, I want to again welcome you here today. I hope you will have a productive day and an interesting one. Thanks for coming.

CHAIRMAN BERG: Thank you. We'll turn now to Bob Cook-Deegan who wants to bring us up to date on some of the current developments and the status of this discussion before we begin our own. Bob.

MR. COOK-DEEGAN: My name is Bob Cook-Deegan, and I am the Project Director for the OTA's project on mapping the human genome. We started last December at the behest of the House Committee on Energy and Commerce, and we've since gotten requests from the Senate Committee on Labor and Human Resources and the House Committee on Natural Resources Subcommittee of the House Committee on Science and Technology.

Taking apart the jargon, essentially what we've got is the committees that authorize spending for the NIH and DOE in the House and the Senate who have requested this study.

There are four principle issues that we have to deal with in the OTA study. The first of these is funding levels of whichever components of projects there are. Second is how do you organize the efforts. Third is how do you impart scientific and industrial advice to those efforts; and a fourth level of issues is international issues, international cooperation and economic competition, and how does this set of projects plus into that.

The topic today deals only with the first of those four issues, and that is the funding levels. The reason we have decided to convene the workshop which was not initially planned, in fact, in the project plan, was it's becoming clear, especially in the last month, that the legislative pace has picked up quite a bit. Our original plans were to have the document available for the authorization and appropriation cycle next year, which would set up the 1989 budget.

It appears likely that several bills -- Well, there are actually several bills already in the hopper, and there are likely to be a few more by the end of the year. And it's quite possible that action would be taken by the end of the year, this calendar year.

That being, the case, we kind of had to accelerate our process, and one of the most important things we need to figure out is what's going to be in the bill. We, obviously, are not going to determine that, but I would like to be able to provide a menu to the Congressional committees that says these are the components and these are the things you need to think about as you're preparing a bill.

We fell upon the idea of doing a workshop on costs really because I was trying to write the opening chapter and delegated the responsibility for figuring out cost projects to Dave Guston; and we struggled with this for many, many weeks trying to come up with a way of figuring out what we should say about them. When you can't figure out an answer, you quite often fall back to a process, and this is the result of that decision.

The purpose of today's workshop is threefold. One is to inform legislation. I think that's the most important purpose. That is to say what's going to be in, what's going to be out, what are the various components and what do they cost.

A second is to -- in discussing budgets, to figure out what are the various component parts. Define what it is that people mean by human genome initiative. Thirdly, it's to help us write the OTA report.

Very briefly, let me discuss some of the materials that are in your packets, for those of you who are on the panel. This is in part to help us structure the discussion for the rest of the day, and in part to help guide discussion on costs in a general sense.

What i'm going to do is go through the history of the cost projections that we know of. The first of these was done by the University of Santa Cruz, three faculty members at the University of Santa Cruz, in late 1983 or early 1984 fell upon the idea of constructing a physical map and developing sequencing technology, a two-pronged attack on providing a tool for biomedical research in the United States. Those are still the main two components of what people are talking about as a human genome project.

They made cost projections in a memo that they sent to Professor Sinsheimer who was the Chancellor of the University of Santa Cruz -- still is. That memo is in your packet, and that was, to our knowledge, the first time cost projections were made.

I won't go over the details of what's in them or funding levels, but they're there for your information.

The next significant step was the projections that Christian Burks made for the 1986 Department of Energy meeting in Santa Fe, the follow-up to that meeting. A third set of cost figures, and the most publicly available one, is in the HERAC Subcommittee report that was released in April. Those are the figures that have been widely reported in the press that would start at $40 million in 1989, increase up to $200 million per year by 1993 until $1 billion has been expended.

That report is, to some degree, silent on who should spend that money and exactly what it should be spent on, but the one key component on which those projections were based were the handwritten notes of Lee Hood that are also in your packet.

That set of projections was done at a meeting in January, or was it February, in Denver; and that's there for your information.

A fourth exercise has been undertaken by the National Research Council. They have written a report, a very distinguished panel, some representatives of which are on this panel in turn. That report is being expeditiously reviewed, but is not now available; and I would guess it will be in the next couple of months, but I don't really know what the details of that are.

We will include analysis of that when we do the final OTA report, as soon as it's available. We'll analyze those costs. We wanted to have an independent and highly conspicuous process for looking at our cost projections, however.

Then there are all sorts of miscellaneous projections that have been cited, a lot of them cited to individuals in this room. You can defend or retreat from your own projections, as you see fit over the course of the day.

The overall context for this meeting, I think, is best characterized by a statement that Lee Hood made at the Denver meeting. That committee had been struggling -- This is a two day workshop , and the first day there was kind of a struggle to figure out exactly what was going to be in the recommendations of the HERAC Subcommittee. Lee Hood that morning came in and kind of turned things around by making the following statement.

He said, essentially, should the budget be driving our vision of what we include in this, or should the vision drive the budget? What I would like to do is define a slightly different process. I would like to define the vision of what people are talking about, but my own experience in Washington suggests that we can put the whole vision out there and Congress and the Executive agencies will fund those sections of it, those pixels of it that they find most relevant to their missions.

So I would like to, in using -- In a sense, I'd like to use the budget discussion as a proxy for figuring out what belongs in these projects, and to the degree that we can, figure out the priorities. I don't think we can come up with comprehensive strategy, but to the extent we can do so, you'll help me writing my report, and I think you will also help everybody in thinking through what might be in these projects.

The final comment on process: There will be a transcript of this meeting available in a couple of weeks and, based on that, we will be incorporating that into the OTA report; and Christian Burks has agreed to do a short write-up with the main conclusions of the proceedings today. Of course, OTA will be independently doing the same thing internally. So we'll have kind of two sets and a transcript available at some point.

Our report is scheduled to go to our Technology Assessment Board, our Congressional board, in December of this year, and it would therefore be released sometime next Spring. As I say, there may be legislation while our report is still being written, however; and to the extent that we can, we will be informally participating in Congressional briefings and things like that.

So whatever information is follow-up to this workshop, if an idea hits people either in the audience or those of you who are around the table, something that was forgotten, some important component that we should be thinking about, please send us letters or call us on the phone. As always, we'll be trying to gather as much information as we can and then condensing that into the OTA report, and also into the OTA briefings for Congress and anybody else that asks us to brief them.

Thank you.

CHAIRMAN BERG: Thank you, Bob. I'm struck by the opening comments. on the one hand, the justification for the OTA which was to advise the Congress and government agencies when they want to make new policy, and you're telling us now they want to make policy and can't wait until the advice, which seems ludicrous; but I suppose that's the way they operate.

I have to apologize for me having to sit in this chair, because I have no business being here, and I tried to make that point to Bob and the other organizers of this, because -- Perhaps they asked me to do this, because I probably had the least involvement of anybody here in this project, at least personal involvement or even professional involvement. I think my one experience with this whole idea was an abortive attempt to try and lead a discussion at a Cold Spring Harbor meeting last year when Jim scheduled -- Jim Watson scheduled this topic for discussion, and Wally and I tried to lead the discussion but could hardly get to the science for -- because of the concerns that -- the ominous concerns that everybody seemed to see in this particular project.

In any case, it's a pleasure to be here where I think everybody agrees that this is a worthwhile project, and now we can get on to talking about how one might go about it in the msot cost effective way and scientifically effective way.

I don't have anymore remarks to make, and so I think let's go to the first item on the agenda which is -- Okay, I'm sorry. For identification purposes, we'll have everybody go around the room and just mention their name and their affiliation, and then we'll start.

MR. HERDMAN: I'm Roger Herdman, Assistant

Director of OTA for Health and Life Sciences.

MR. PEARSON: I'm Mark Pearson, Director of Molecular Biology and Central Research at the duPont Company.

MR. CARRANO: I'm Tony carrano. I'm the Genetics Group Leader at Lawrence Livermore National Laboratory.

MR. SULSTON: John Sulston from the MRC Laboratory of Molecular Biology in Cambridge, England.

MS. HOBEN: Patricia Hoben, Analyst from OTA working on this study and a biotechnology study.

MR. WATSON: Jim Watson from the Cold Spring Harbor Laboratory.

MS. DONIS-KELLER: Helen Donis-Keller, Senior Research Director in the Human Genetics Department at Collaborative Research.

MR. CAHILL: George Cahill, Howard Hughes Medical Institute.

MR. GILBERT: Walter Gilbert from Harvard.

MR. HOOD: Lee Hood from Cal Tech.

MS. COURTEAU: Jacqueline Courteau, Research Assistant for this project at OTA.

MR. McKENNEY: Keith McKenney from the National Bureau of Standards.

MR. BURKS: Christian Burks from Los Alamos National Laboratory.

MR. STEVENSON: Bob Stevenson, American Type Culture Collection.

MR. GUSTON: Dave Guston, Research Assistant, OTA on the Human Genome Project.

MR. BERG: I'm Paul Berg from Stanford.

The first item on our agenda -- I should point out, as Bob mentioned already, the purpose of this meeting is to see if one can arrive at some price tag for some of these projects. In trying to think about this before I came, I realized how handicapped I was because I hadn't done any of these, and my own lab isn't involved in any of them. So very difficult it is for me, and I don't know whether it's going to be as difficult for those of you who are right in the middle of trying to do much of this. But in any case, I suppose we have to approach this in a way as if we were writing a grant proposal, and you have to put some numbers down on your grant proposals. So we'll take some shots at it in the same way. And for those people who review our grants, they will now look in on how we arrive at these magic numbers.

I think it would be useful in the RFLP map estimate if, Helen, you could give us a little sort of exposition and summary of where you think this project stands at the moment and what your expectations are for the time it would take to arrive at a 10 centimorgan map, and I think that should be the first target, to reverse those two, and ultimately to something with a resolution of 1 centimorgan map. I'm thinking of a genetic map, the RFLP map.

MS. DONIS-KELLER: Okay. First, let me just start with a couple words on why it's important and how it relates to the sequencing that will be done, the creation of a physical map.

Well, first of all, a genetic map will allow people to focus on particular regions that are of interest. For example, there are genes that have been localized to chromosomes, two genes for manic depression on the X chromosome and chromosome 11, a gene for neurofibromitosis on chromosome 17 region, and a region on the small arm of chromosome 16.

It would be really useful if, once we have located these genetic regions, and I'm talking about regions of about, say, 10 base pairs, 5 to 10 million base pairs -- It would be really useful at this stage then if one were able to go into the refrigerator and pull out a set of overlapping clones that covers that region, sequence that region and then look at all the possible coding sequences there to see what is the gene that causes manic depression, and try to sort that out.

So what the genetic map can do is allow us to focus on regions that are of extreme usefulness and importance first in the sequencing project.

The other thing that the genetic map can do would be to help align physical -- the physical map. That is, say, you have an overlapping set of contiguous clones. You don't know which end is up or which end is on what part of a particular chromosome that you're working on. The RFLP map can help you do that. It can tell you which direction the ends of the clones are pointing.

So from those points of view the genetic map will be very useful, and the physical map. Now as far as calculating the cost of the 10 centimorgan and 1 centimorgan map, I've actually prepared a little talk on a 5 centimorgan map, because (Laughter) -- Okay. We've been talking this over, and actually my group has been working on this for several years, and our conclusion is that a 5 centimorgan map would really be a useful map in terms of relating it to the physical map, and would be an extremely useful map from a genetic point of view.

A 10 centimorgan map, of course, would be useful, but it doesn't give us the kind of resolution we need in order to tie that into a physical map, even if there are advances in the technology so that one could do megabase cloning.

A 1 centimorgan map, from our personal experience, is that it's just not worth it to make a genetic map at this kind of resolution. You need too many family resources. The amount of work that goes into that compared to making a 5 or 10 centimorgan map -- it's just very, very much larger. So I would like to talk about a 5 centimorgan map.

Now where is the world in terms of making any sort of genetic map at all? Right now there's not a human genetic map, a primary map, that is with no gaps of, say, 40 centimorgans. When I say 40 centimorgans, a centimorgan is a unit of genetic measure, and it's thought to be about equivalent to about a million base pairs.

There are several groups that have been working on creating a genetic map, and I can tell you that a map of about a little over 300 RFLP markers has been worked on in our group, and we've now mapped several hundred markers back to chromosomes. If you look at the size of the human genome, which is about 3500 centimorgans, and divide that by, say, 300 markers, then you have a map that's about the resolution of 10 or 11 centimorgans, something like that.

We've been doing this project, and we've mapped these 250 markers back to chromosomes, and that would give us on average markers about 15 centimorgans. I can show you one example of one of the chromosomes that we've done there.

So another group is also working on this project and has several hundred markers as well. Then there is an international collaboration called The Center for the Study of Polymorphism that's been established by Jean Dausset, and we're all using the same set of reference families. So in time a complete connected map of the entire genome will be created and shared -- the information shared with all researchers throughout the world.

My estimate as far as how long that will take in terms of what's being done now and what I see in the future is that it's still several years away; but very good progress has been made.

So what I'd like to do then is just run through what I think the calculations are for a 5 centimorgan map. Okay?

CHAIRMAN BERG: Fine. I'm going to reemphasize something I think you said, and I remember it stunned me and amazed me when you said it two weeks-- several weeks ago in Palo Alto. That is, your estimate is that you're likely to have that 5 centimorgan resolution map within the next two or three years, maybe even less, one to three years, I think you said.

MS. DONIS-KELLER: I think the world community of geneticists will have something like that put together in a few years' time.

CHAIRMAN BERG: Right. So in effect it's almost a done job. I said almost. I didn't say it is. I said it's almost a done job.

MS. DONIS-KELLER: Well, as you know, that depends on the support, the financial support that people get from the private research dollars that are available as well as government funding and private institutions; and if funding continues, then, yes, that will happen.

CHAIRMAN BERG: Okay. The reason I emphasize it is because I think that was not in the mind of many people that I've talked to before who have raised that as the highest priority, as if it was the major next step to do and was going to require a lot of funding. I think, when I heard you, I came to the impression that it was very close to being at least at the 5 centimorgan level, and you're arguing that that's all you need.

MR. GILBERT: Paul, why don't you wait until she gives the numbers, and then have this discussion?

MR. HOOD: May I ask for one clarification?

CHAIRMAN BERG: Sure. Go ahead.

MR. HOOD: I wanted to raise the question, Helen, do you see any significant technological advances in being able to do this kind of mapping? Is that -- Are those advances calculated into your cost, or do you -- Are you thinking entirely in terms of the technology we have today?

MS. DONIS-KELLER: In terms of building the RFLP map I'm looking entirely in terms of the technology of today. As far as being able to use an RFLP map of the resolution of 5 centimorgans, relating that to the physical map, I'm relying on technology improving so that one can actually make a set of overlapping clones that could cover an entire chromosome; and what I'm relying on at the moment is the megabase cloning, for example, the technology that Maynard Olson has pioneered, cloning large segments of human DNA in yeast will actually become available.

CHAIRMAN BERG: Why don't we go to your estimate of the cost, and then we'll come back to --

MR. HOOD: Let me just make one point. It is very conceivable to me that there will be real automation of techniques for doing this RFLP mapping that could really facilitate in striking ways, and maybe we'll hear a little bit later about some of those kinds of things. So we ought to keep that in mind as we make these kinds of decisions.

CHAIRMAN BERG: Bob.

MR. COOK-DEEGAN: I wanted to get one clarification also. When you're saying 5 centimorgan, I gather what you're saying is average, 5 centimorgan?

MS. DONIS-KELLER: On average, right.

MR. COOK-DEEGAN: Do you know what -- Are there any mathematical projections that would tell you what would be the largest gap, if that were the average coverage?

MS. DONIS-KELLER: I haven't done those calculations, but once you find out where the gaps are, they can be filled in by going to chromosome libraries and that sort of thing. So right at the moment, I don't see what those problems are. In looking at the map that we worked on, we can see where some rather large gaps are.

CHAIRMAN BERG: George?

MR. CAHILL: If I'm correct, there are rather hot and cold spots for recombination. So a centimorgan can be an order or two or magnitude difference in length going anywhere from 10,000 or 20,000 bases up to ten million. So how does this impact on what you're doing?

MS. DONIS-KELLER: Okay. There is a -- It's thought that a centimorgan is about a million bases, but in fact there are recombination differences between males and females that in males, in a particular interval between two markers there may be less or more recombination than in females. So that gives you two different pictures of what the genetic distance is.

When you try to translate that to physical distances, you don't really know, is it really 10 million bases or is it 1 million bases. We don't know enough from the mapping that's been done with genome to know how many of these places which some people call hot spots for recombination there are; but it's certainly true that it is somewhat ambiguous.

CHAIRMAN BERG: You want to then go on to your estimate. We'll come back to this more general discussion.

MS. DONIS-KELLER: I guess I should point out that Dr. Cahill here represents another group in Utah that's been working on the RFLP map, too. So he may have some comments as I go through this.

CHAIRMAN BERG: What would you say? How many laboratories whose principle target is on this kind of mapping? Four or five?

MS. DONIS-KELLER: Two.

CHAIRMAN BERG: Two.

MS. DONIS-KELLER: I think there are two laboratories in this country where they are making a really serious attempt to put together an RFLP map. Other laboratories in the country that work on RFLP maps are usually interested in a particular chromosome or a particular set of disorders.

Excuse the misspelling on the first sentence there. The question is then, how many RFLPs do you need to map the human genome. The human genome is about -- The autosomal genome is about 3300 centimorgans, and then we have the X and the Y chromosome. There's only really a recombination between the X chromosomes and the Y chromosome can pretty much be neglected for the purposes of genetic mapping; but altogether, it's about 3500 centimorgans.

We just talked about a unit of genetic measure was centimorgan. For purposes now of working, we consider 1,000.

Here's what that looks like in terms of the chromosome. Some are much larger than other chromosomes, and this just shows you the set of human chromosomes 1 through 22, and then the sex chromosome. The largest chromosome here, No. 1, is about 8 percent of the human genome.

So what we're talking about in making the RFLP map is a set of evenly spaced genetic markers that cover all of these chromosomes, so that we can then use this genetic map as a tool to identify various genetic traits. Some of those traits are diseases. What we plan to do, and have been doing, is to locate a particulate disease gene, say, between two genetic markers along these chromosomes.

Now I said a map of about 5 centimorgan resolution on average was the first target to really aim for. So if you simply divide the size of the human genome by 5 centimorgans, you come out with about 700 RFLPs are required to make this map.

Now something else that's important, and that is that the RFLPs must be informative. By that, we mean that you're going to use these genetic markers. It's important to be able to trace the inheritance of the markers in families, and that means that the RFLPs have to have a number of different variant forms or alleles. So that quality -- how good -- A genetic marker is quantitated by a term called PIC, polymorphism information content, or heterozygocity, the frequency of heterozygotes

So perfect RFLPs have a PIC of one and a heterozygocity of 100 percent. Now that would be ideal. However, these highly polymorphic loci are rare. They are hard to find, and we've searched very hard for them. So one can't really depend on coming up with 700 of these very fine markers in a short period of time with today's technology; but in fact one really doesn't really have to have perfect genetic markers in order to build a useful genetic map.

So based on our experience, I would say that a more practical distribution of informing us of these markers is that about two-thirds of them have the information content, the PIC value between .3 and .5, and about a third of the markers would have information contents above .5. And that set of markers will be useful for creating a genetic map that can be used.

MR. HOOD: Helen, how is PIC defined?

MS. DONIS-KELLER: PIC is the probability that, in any given offspring one will be able to identify the parental alleles.

Now in costing this I've just used a rough calculation that says that one full time equivalent or one full time person costs about $100,000 per year; and, of course, that varies, depending on your overhead and so on and so forth, but I just used a rough number of about $100,000.

Okay. So to make this genetic map then, one needs a set of RFLPs that would yield polymorphism. One also needs a set of families in order to do these inheritance studies to put the genetic map together, and then linkage analysis techniques.

I'm going to talk a little bit about the RFLPs, but just to get through the families, this is the kind of family structure that one needs to do the genetic mapping. We have four grandparents here, the parents and a large sibship, a large number of offspring. We refer to these in genetics as meiosis. There is a reference set of families that have been collected -- and that's what the study of polymorphism is all about. Those families have been contributed by various researchers and comprise this reference set of forty families. So in fact a set of forty families that's needed to make the genetic map is in place and the DNA is potentially available. This is what we have been using ourselves, as well as drawing up some cell lines ourselves for these families.

Then the genetic resources that are needed are about 700 RFLPs and forty reference families which the average sibship size amounts to about 640 meiosis that can be used for the linkage studies. That's sufficient to make a genetic map of a resolution of on average 5 centimorgan. And linkage analysis techniques are being developed and are multilocus analysis techniques, which are really critical for the larage number of RFLPs.

Cell programs have been developed are in place.

Just to go through then what is necessary to grow the RFLP map: A previous study that we've done in our lab identified 250 mappable RFLPs with five personyears of effort. One thing to note is that it will become increasingly difficult as time goes on to find yet more new mappable RFLPs. So I filled in this difficulty factor.

This is assuming a certain degree of technology improvements but nothing really substantial. So it's really an estimate based on today's technology and how difficult it may be to reach a number of 700 RFLPs from our initial starting point of 250 RFLPs.

So if 250 represents five personyears of effort, then the next 250, I would say, would be about eight personyears effort, and the final 200 RFLPs in this particular project would be twice as difficult as the starting first 250. So that would be ten personyears of effort. So the total then is twenty-three personyears of effort.

I would say, in the academic community, considering, as I said, the usefulness of the polymorphism, that we are again only at around 250 to possibly 500

RFLPs that can then lead to mapping, but I don't think that there are 700 RFLPs right now, at least according to what's been published responsive to the 700 needed for this RFLP map.

Many of the RFLPs that have been described are not usefully polymorphic. The distribution of alleles is very one-sided, so they're just not that informative.

So the work involved in doing this then, so you can see what the efforts are based on, is that when you identify RFLPs you construct and screen chromosome libraries, genomic libraries from bacteriophages or cosmids or simple chromosome library. Then this just tells you the typical screening pattern when you test for polymosphism consists of five unrelated individuals.

So then you test these clones. We're talking DNA clones, -- to see whether or not they can reveal polymorphism in human genome. What you also do at the same time is create an archive collection, which is frozen down so that you can always go back to that archive collection, and you maintain working stocks.

Then a characterization of these RFLPs that you have identified from the screening study involves testing more individuals so that you can get a real picture of how usefully polymorphous the RFLP is. You score the alleles, determine length and so on and so forth; confirm that they represent single loci. It's very critical in the genetic map that you're not looking simultaneously at two loci. You need to have one reference point that you're using as your genetic reference point for each RFLP.

Then there are some improvements that need to be done along the way, and that takes time. Then one needs to establish a data base and enter that information into the data base, check it to make sure it's correct. And all of that takes time as well.

So for this effort, the characterization part, my estimate is that it will take about twenty personyears to create this resource of 700 RFLPs that can be then used in the genetic mapping. And just as a guide, I think one person can accomplish this characterization on about 35 RFLPs in a year's time.

CHAIRMAN BERG: I'm just confused here. Was that a new estimate, or is that to be added onto the estimate that you gave earlier?

MS. DONIS-KELLER: It gets added on. I'll have a grand total at the end. So the identification of the first set of RFLPs is 23 personyears, and then you add on the characterization, preparing these for linkage analysis is another twenty personyears.

So the linkage mapping process -- The work involved here is the maintaining of the lymphoblast cell lines from which you purify the DNA to do linkage studies with. And I estimate to make DNA available so that 700 RFLPs could be mapped, this is about a ten personyear effort.

Then seven blots need to be done, first checking parents for informativeness and then following up on families, the set of reference families. And I've just written out one estimate here. Say, if you have twenty blots times two enzymes times 700 RFLPs, that's about 28,000 blot uses that are necessary. So this is a fair amount of labor that goes into this.

Okay. So I estimate that that part of it, doing the generation of the data that's been analyzed, is about a thirty personyear effort. Then the linkage calculations: There is program developments that need to be done, data base management, and the actual generation of the male and female maps; error checking and data review, I think, constitutes about fifteen personyears of effort. The error checking is very critical. It can be very time consuming. One sometimes needs to go all the way back to the initial films, the linkage mapping films, to check to see have the grandparents been reversed, have the data been entered incorrectly on the first sheet which then gets entered back into the computer. And there are various computer programs that flag regions of chromosomes that tell you to go back and look at this region of the chromosomes, that there are some problems here. That's a lot of time.

Then checking to make sure that you have complete coverage of the entire genome is done by a variety of methods, but one method is to physical localizations. You take probes and you hybridize them back to the cytological map, the physical map of the various chromosomes. It's not a physical map in the sense that you can juse take those clones and use them. It's a physical map in the sense that the chromosomes are stained, and you know which regions, where the tips of the chromosomes are, where the center is; and you go back and make that physical localization with one of your RFLPs to say at the very end of my map am I at the tip of the chromosome. I'm calling that cytological orientation.

That effort for 700 RFLPs, I think, is about a ten personyear commitment.

So just to add it all up then: The screening is about 23 personyears. The characterization, 20. So that's about forty-three -- 40 percent of the effort here. Then the linkage mapping part is 65 personyeras. So that gives us a grand total of 108 personyears times the $100,000 per full time unit. It comes about to about $10.8 million which, in the context that we're talking, is really spare change. And I'd like some of it.

There are other costs that are not in this calculation. There is a capital equipment cost, and that can vary a lot, depending on the number of people you have using the capital equipment resources. I've just roughly estimated that to do this with a group of, say, fifty people working at one time, it may be around a $5 million investment, just based on what a centrifuge costs today, the computer equipment, etc. But it depends also on how this kind of RFLP mapping work is done. If there are a lot of various places doing it, then the communication support between those various locations needs to be very substantial, versus if one team is working in one facility. So I haven't put an estimate on what that would cost.

Then, of course, the facilities, the physical facilities that need to be built and maintained. So I really don't have a good figure on what that would cost, but maybe you would say another 10 million at the most. I don't know.

So that's what I think it will cost to make a map of this kind of resolution. I should stress, that means that on average it's a 5 centimorgan map. There will be regions in which there are larager gaps that need to be filled in. That can be done using today's technology by going to particular chromosome libraries, screening for RFLPs from those particular libraries and filling in the holes that way.

I think I have -- I'll show you one chromosome example here. As I said, 250 of our RFLPs and mapped those back to chromosomes, and we've since added more RFLPs to that, but this one chromosome was generated from this study. In these RFLPs we didn't know what chromosomes they were on in the beginning. We made a linkage through and then tacked that to a chromosome by doing physical localization of, say, one of those probes in the linkage group.

So you can see here the male and female chromosomes. The male chromosome we have here 207 centimorgans total, and the female chromosome is about 299 centimorgan. So the female chromosome appears genetically larger than the male chromosome.

Then you can see that it is one continuous linkage map of the entire chromosome, but there are regions that are larger than the 15 centimorgans that one would have predicted on average the distance between the markers, just based on the number of markers divided into the size of the genome. But it is a connected map, and these regions can be -- say, this region here, 157 to 207 centimorgan. There are about 50 centimorgans that can be filled in, and there are markers over here which have not been placed within some intervals, and there are sets of markers.

So once you have, you know, a map of this kind of resolution, you can go back in and make it a 5 cM distance between each marker for all of the markers, so that you do have in fact a 5 cM resolution map for all of the chromosomes. But again, my estimate was based on making this on average 5 cM map. It does not make an estimate for guarantying that you have 5 cM -- no more than 5 cM between any two markers, and that would be a certain amount of work more than what I've showed you. But it's doable, and it is not significantly -- It's not a significant effort. I don't think it would be that difficult. It's just simply more work.

CHAIRMAN BERG: Based on that hundred-and-some-odd manyears, how did you come up with the estimate then, roughly in the next one to three years or two to three years? You only have 25 people that have been working on this, and there's another group at Howard Hughes?

MR. CAHILL: About the same number.

CHAIRMAN BERG: About the same number. It's two years. Okay.

MR. STEVENSON: What is the end result that you desire for this, simply to describe on the map what these various loci are and what their significance is, or is there a further utility in terms of having this described and having these pieces of material available for others?

MS. DONIS-KELLER: The purpose of building the map is to make a biological tool that one can use then to identify all the various traits that combine themselves to make a human organism. What does it take to form an arm or a hand, or what does it take to make hair color or behavior patterns and all those kinds of things. With an entire map of the human genome, you can use that to go and simultaneously look for a combination of genes which can cause a particular complex of phenotype.

So the map is really a tool, and this is how I see using the tool for what I'm interested in and what many geneticists are interested in. Other people may be interested in particular regions of chromosomes to look at. Other people may be interested in just understanding what is it that makes up a genome. Why are there differences between males and females in recombination. There are a lot of basic research questions that are of interest that will apply to a whole variety of organisms.

MR. STEVENSON: How are those questions going to be studied, using what model systems?

MS. DONIS-KELLER: Well, you've asked a very large question.

MR. STEVENSON: The point is, it's one thing to go out and map new terrain and say, okay, I'm going to call this Mt. Victoria or Mt. Everest or whatever and so forth, and just give a descriptive type of thing. It's something else to mine that territory, to farm it, to do something with it in terms of the application of that to some human end or utilization.

It's one thing to map this, but if everybody is going to study one particular centimorgan piece, then you have to assume that you have some process whereby you generate duplicates of that centimorgan piece and you make it available to other people to work on.

MS. DONIS-KELLER: What you're talking about now is the physical clones themselves, and that gets into the area of making the overlapping set of physical clones; and that's what the other people here have the expertise in. However, you use the genetic map as a tool, a biological tool, in exactly the same way you can use it to make the map in the first place.

You have a set of families that have a particular trait. They are segregating a trait that you're interested in studying. So you take that set of the RFLPs that represents the genome, apply it to that set of families, and then you're able to home in on what particular region of the chromosome -- genome is responsible for that trait that you're following. Then you go and you look at that region physically in the genome and identify what the genes are in that interval that could cause the trait you're interested in.

MR. STEVENSON: Okay. But your estimate of cost does or does not include making sufficient amounts of that material to distribute? Or it is the first cut at describing where it is?

MS. DONIS-KELLER: What I've described is what it will take to make a genetic map of this resolution, what it will take to make the information available so that it can be used. It does not take into account making a whole lot of copies of these various RFLPs that can be distributed. Again, that's something that would need to be calculated.

CHAIRMAN BERG: Jim?

MR. WATSON: Helen, two questions. One is what would the cost be if you did the 1 centimorgan map? And, second, the National Academy committee and Maynard also has said, he believes that you really needed the 1 if you were going to link together.

MS. DONIS-KELLER: That's right.

MR. WATSON: Now you disagree with him?

MS. DONIS-KELLER: Okay. The reason I'm enthusiastic about a 5 centimorgan map is precisely because of the work that Maynard Olson is doing. If you have a 5 centimorgan map -- Say, you have a map in which the spacing is no more than 5 centimorgans between two markers. That means if you're doing megaphase cloning and you have overlapping clones of, say, a million to two million bases, you can easily make the overlap between those intervals hybridize back to, say, an overlapping set, and you have the interval identified for what you want to do in sequence.

MR. WATSON: Now at that meeting Maynard knew what he could do, and still he said he needed 1. So was he not believing his own work, or do you have a fundamental difference?

MS. DONIS-KELLER: Okay. Some people are more optimistic about the technology that Maynard is developing to do megabase -- At that meeting, he was more pessimistic about the ability for one to make these large overlapping clones. At the meeting the largest clone he had identified was 900 kilobases. Okay? Even if you don't have that, I suppose -- Well, people can tell you how reasonable it is that one can develop an overlapping map of an entire chromosome without the megabase cloning. Okay?

CHAIRMAN BERG: Jim, as I understand -- There are two purposes. Maybe you need a 1 megabase resolution in order to use it for the physical mapping, organizing the cosmid clones; but for genetic mapping or for mapping human diseases, then the 5 should --

MR. WATSON: But -- looking at it from the viewpoint of going out to sequence --

CHAIRMAN BERG: No, I understand. I just want to make that distinction.

MS. DONIS-KELLER: I maintain that a 5 centimorgan map will give you the kind of resolution you need to relate that to a physical map that you use as a tool for the physical map. The 1 centimorgan map -- I've done some rough calculations just based on our experience in the lab. It's just so much more effort and it will take so much more time that it's not worth doing when a 5 centimorgan map is going to be okay; because the distance between two 5 centimorgan intervals is only 2.5 centimorgans. People are now doing physical mapping of various intervals on chromosomes, and I think that is possible,within the realm of possibility.

CHAIRMAN BERG: May I ask George, who I think is the other major partner in funding, Federally funding, an effort along these lines, if you could give us a sense of where your thinking is?

MR. CAHILL: I'm really just echoing Ray White and Joe Markle in Utah which were supported by Hughes, and they're putting about $2 million a year with the same staff, equivalent to what Dr. Donis-Keller just presented. But their feeling is that a 1 centimorgan map will probably bve needed, particularly -- And they're going to aim in this direction. I can't quote any further.

CHAIRMAN BERG: Excepting all the concerns?

MR. CAHILL: Excepting the -- and particularly because this variability and recombination, that if one is interested in an area -- And also there's a practicality here which Dr. Donis-Keller didn't get into, namely, that for genetic -- clinical genetic use, the 5 centimorgan gives you too much error to make good genetic assigns. You've got to get closer to the gene of interest. Therefore, eventually, you're going to have to clone down further, whatever, when you have a disease situation.

MS. DONIS-KELLER: But that's just on focus regions; and in fact, the way the physical work will be going, the reason for having markers that close is so that one can do linkage diagnostic tests. I think those will be superceded at some point as soon as the genes are identified, and direct tests will be developed.

So making the 1 centimorgan map is extremely difficult, because you rely -- When you are so close -- It's very hard to know how close you are when you are so close, when you're within a couple of zone markings.

An example of that is looking at the cystic fibrosis locus. It's now been more than two years since the initial identification of markers that are linked to CS, and it's still ambiguous as far as what is the order of the markers within, say, 5 centimorgans of that locus. A lot of work has gone into that. You need doubly informative markers in order to detect a crossover within that interval. You need large family resources.

So I just don't think it's practical to make as your goal a 1 centimorgan map when a 5 --

MR. GILBERT: I'd like to move that we're getting too much into a question of what ideal policy should be. I think it's simpler to sort of restrict the conversation to what are the actual goals.

CHAIRMAN BERG: Wally, can you speak up?

MR. GILBERT: I think it's simpler to restrict the conversation to what are the actual goals and what are the actual estimates. George just said Hughes' goal was in fact a 1 centimorgan map.

MR. CAHILL: We're pushing in that direction.

MR. GILBERT: That's right. Do you have any estimate from inside Hughes as to what you think that's going to cost?

MR. CAHILL: No.

MR. GILBERT: Or how long it's going to take?

MR. CAHILL: It's speculation now, because of the newer technologies, and also looking at these highly variable tandem or piece, the kind of jeopardy fragments, whatever you want to call them. There are other technologies which, just like primary sequencing, is going to move this very rapidly; and that's part of the equation of going into the 1 centimorgan map, is capitalizing. Optimistically, there may be better technologies to look at these variations.

CHAIRMAN BERG: What are the better technologies? I'm not clear on that.

MR. CAHILL: Well, Helen, you're more qualified to speak on this than I am.

MS. DONIS-KELLER: Well, there are some set of RFLPs that have been described called mini-satellites.

White's group calls them variable tandem repeat elements. Those mini-satellites detect a large number of loci in the genome. So in order to make thsoe useful, you have to do cloning so that you can isolate the one that represents a particular chromosome.

I think that's a fair amount of effort, and I see that as a way to identify potentially usefully polymorphic loci; but I don't see that as a shortcut in terms of what it takes to make that useful to single locus probe. But, you know, I really think it's probably not relevant to get into an argument, is a 1 centimorgan map the goal or a 5 centimorgan map.

When you get to a 5 centimorgan map, I think it will become obvious that's it's useful to tool, and it's the eventual goal as a 1 centimorgan map, all the better.

MR. GILBERT: What is the -- You're finding then that the goal of collaborative research effort will in fact be the 5 centimorgan map overall.

MS. DONIS-KELLER: Well, our goal at the moment is to have a map that's genetically useful as a tool, and as soon as we have something or some regions of chromosomes that are useful, we'll be able to use those as a tool. I personally think, discussing this with people that I work with, that a 5 centimorgan map is a very useful tool to use for all of the purposes that we've just discussed, and particularly useful for the physical mapping purposes.

CHAIRMAN BERG: Mark, you wanted to say something?

MR. PEARSON: Helen, it would useful, I think, for Congress, as they're going to be one of the digestors of this report as well sa the agencies proper, to have some sense of the variation you see in your current estimate, since we all know these are guesstimates at this point in time, and also to have some sense at least of order of magnitude, in your view, that a 1 centimorgan map might cost.

MS. DONIS-KELLER: You're the ones that have the goal.

MR. CAHILL: I don't want our trustees to know.

MR. PEARSON: The dilemma, of course, is that, if we don't go ahead and at least give some sense of range here, Congress is going to be left in the situation of assuming that "N" dollars spent on "this project" will generate the result; and if that's off by a factor of ten, I think we ought to say it up front.

MS. DONIS-KELLER: That means you're looking at 3500 RFLP markers minimally, and then you have to guaranty that they are no more than 1 centimorgan apart from each other. I'd say, orders of magnitude more. It's extremely expensive project, and I could go back and make more calculations; but I don't think it's useful.

MR. COOK-DEEGAN: So you would have to make a minimum of one order of magnitude at least difference if you went to the 1 centimorgan?

MS. DONIS-KELLER: What would you say?

MR. CAHILL: Well, there's a -- Again, a question here that's different. There are about 2,000 or 3,000 markers already been assigned to loci which will help in the contigs and all of that. You don't need the RFLP variation to do your contig mapping. We're talking about RFLPs, highly variable things which are best used for genetic reasons. Am I correct on that? And, therefore, to go for the 1 centimorgan has more of a clinical use than the need for the contig, lining up.

Now I'm speaking beyond the limit of my knowledge, with you pros around here.

CHAIRMAN BERG: I thought it was the other way around, that the 5 centimorgan map would be for clinical genetic --

MR. CAHILL: Well, it's both. Both.

CHAIRMAN BERG: Except for the exception that the likelihood of being correct is, obviously --

MR. CAHILL: What I'm trying to infer, the polymorphism is not critical for the primary ordering of fragments. A simple DNA probe will do it as long as it's got a single assignment in the human genome. That's the point I'm making.

MS. DONIS-KELLER: You can use just single fragments of DNA to orient yourself if you're sure that you know that they map back to a particular chromosome. The mapping back to the chromosome regions is very crude. You can do the physical mapping, and you know it's in one band on a, you know, a cytological map. But, yes, you can take a particular probe and know that it's on the shoreline of chromosome 4, hybridize it to your set of overlapping contigs and say, aha, this set is on the shoreline of chromosome 4; but you cannot take a physical probe and relate it to a genetic probe unless the physical probe has polymorphism.

So if you're suggesting just simply hybridizing back to the physical probes, you also have to have those things spaced in regular intervals, because if you have one gene and then say -- say it's a hundred million bases away. You've got to have an overlap that is that big so that you know the interval.

So, yeah, you can use nonpolymorphic markers, but they have their limitations, too.

MR. HOOD: The other thing I think should be injected here is the 1 centimorgan map might be a lot less awsome because of the technology developments. There will be automated ways for making libraries. There will be automated ways for selecting individual clones and amplifying them up. There will be automated ways for preparing the restriction fragments, and there will be automated ways for doing the mapping and the analysis on five or ten or fifteen different humans. So I think that's a really important factor to calculate in.

We can't make estimates on that. We don't know what the time scale for those is going to be, but we -- There are already automated robots that do some of these procedures very, very nicely. So I think it's a terribly important factor that, in the scale of things here --

CHAIRMAN BERG: Could you just tell us we can't factor it in.

MR. HOOD: You have to take into account that these things are going to change. I can't tell you what the dollar price is going to be, Paul; but it's not going to be two orders of magnitude. I can tell you that. But if you talk in terms of today's technology, I think you're going to miss the point entirely, and you're going to give a false impression that's way too high on how much it's going to cost to map or how much it's going to cost to --

CHAIRMAN BERG: But, Lee, I'm mayby off base on this, but if a 5 centimorgan were sufficient to get on with ordering a cosmid map which would be the next step before you wanted to get a more high resolution sequence of those fragments, why should we be worrying about spending or developing technology for a 1 centimorgan map? I mean, if in fact the 5 is sufficient, I'm not sure I understand why one wants a 1.

MR. WATSON: Maynard says that it isn't. So, I mean, we have a disagreement.

CHAIRMAN BERG: Okay. So we have a -- I just want to get that out.

MR. GILBERT: I also disagree with that. I don't think there's any need for a 1 centimorgan map as a prerequisite for a physical map. I think the only issue on a 1 centimorgan map is if it's going to be useful medically ultimately, and that's where I think it will be useful. I think actually it will be constructed -- my own bet on the future is that we construct it after the cosmid map is in place rather than before.

CHAIRMAN BERG: Yeah. In which case maybe all of that technology that Lee is talking about will be spinning wheels or playing games.

MR. GILBERT: Lee is talking about --

MR. HOOD: That's going to be generally applicable. All of those things I talked about will be used for the physical map, too.

MR. GILBERT: Yeah, but we are talking about automating a lot of common laboratory processes. It doesn't seem to me that the automation saves much money on just doing those processes. And you estimate from Helen's figure five times this cost, $50 million, for the 1 centimorgan map, the only problem being that there is one floating figure in what she used, and that is an estimate that it was harder to find as you go forward.

I don't -- There is -- In a certain sense, it isn't particularly harder to find them as you go forward. They will fall statistically as you go forward, and I would go -- If we had to make an estimate, I would do that figure for the 1 centimorgan map, figuring that in fact if you did this on a larger scale, you would simplify some of the other parts of the procedure, which is Lee's point, and thus bring the cost down on some of the other pieces of Helen's estimate.

CHAIRMAN BERG: What would your estimate be, Wally? I'm trying to --

MR. GILBERT: Well, I'm saying $15 million is an outside figure and, if you want to stretch, you could probably get up to 25. It's not a serious issue.

CHAIRMAN BERG: Given the number of 10 to get the 5 centimorgan.

MR. GILBERT: That's right.

MS. DONIS-KELLER: Yeah. So to get 700 it's at 10 million. So five times that number of markers is five times that amount of money.

MR. GILBERT: You know, if you're thinking markers at arandom, you want to continue picking markers to a total of the 3500 picked at random.

MS. DONIS-KELLER: Right. So on average, you're going to be 1 centimorgan --

MR. GILBERT: That's right.

MS. DONIS-KELLER Ninety-nine percent confidence. So that's many, many --

MR. GILBERT: No, no, no. It's 1 centimorgan on average. You can look at the statistics. There will be some gaps in the statistics of the gaps. So that's not really -- That's not the issue.

CHAIRMAN BERG: Dave?

MR. GUSTON: Helen, you mentioned that data base entry was included in the cost, and what's on the other side of the data base entry? Was that included in your facilities and communications?

MR. COOK-DEEGAN: The long term fixed cost, essentially. After you created the resource, do you have any idea what the components would be that would have to be sustained over a long period of time?

MS. DONIS-KELLER: No, I haven't looked at that at all. Let's say that the estimate on a 5 centimorgan map was me thinking, well, what can we have in the next five years that makes sense that can be related to a physical map, and I think a 5 centimorgan map is going to be a very useful took to move ahead with on the physical map. And if the goal is a 1 centimorgan genetic map, then that will be nice. But I couldn't agree with Wally more that that will come to pass after much of the physical mapping is very well underway.

MR. GILBERT: The timing, for example, that you've exemplified is really that you expect a 5 centimorgan map to be in existence about two years from now.

MS. DONIS-KELLER: Several years from now, provided that kind of labor is put on that problem.

MR. GILBERT: Well, a continuation. That's the continuation of the effort.

CHAIRMAN BERG: You're not dropping your project. So it's going to continue, and the Hughes --

MR. GILBERT: Well, it would be nice if we funded.

CHAIRMAN BERG: Which brings me to one point. We're going to have to move on, but I want to ask a question, because it came up in this recent Science article that highlighted Wally Gilbert and all the other-- Essentially, the issue of the availability of these clones, the markers; and it was very clearly stated that collaboratives policy is not to make those available. Now the question is: If we're talking about mounting an effort in which some other groups who are being funded by the Howard Hughes, and there I suspect all those clones will become available, if some academic institution is funded and they make their clones available, and one subset of them is not available, how is this thing going to --

MS. DONIS-KELLER: You've identified a problem which has really been exaggerated, and it's simply not true that our clones are not available. We are currently sharing our clones with many academic investigators on collaborations, projects that we have a mutual interest in and we share our recources on.

There's no reason to presume that those markers could not be made widely available to the academic community. Right now collaborative has invested a significant amount of its own personal private money to developing these markers. It cannot simply give all the markers away. It's not responsible to the shareholders of the company.

CHAIRMAN BERG: What happens if you --

MS. DONIS-KELLER: There's no reason to assume that that can't happen, that the markers can't --

CHAIRMAN BERG: What happens if you get Federal funding to support the program?

MS. DONIS-KELLER: Then that changes the dynamics of it, certainly. If it's funded by public monies, then, you know, it's a different situation.

MR. GILBERT: I'm sure the government can always buy the probes from Collaborative.

MS. DONIS-KELLER: That's true. I don't think that's a serious barrier at all. It's useful to bring that up.

MR. WATSON: Let's say the government buys them. Will they be available before --?

MS. DONIS-KELLER: As I said, the probes are available to collaborators. We have a large number of collaborators. We intend to increase our list of collaborators.

MR. WATSON: No, but I think you're avoiding the issue.

MS. DONIS-KELLER: I'm not avoiding the issue at all.

MR. WATSON: I mean, will your collaborators be -- I mean, I'm just trying to say, what if Collaborative says they do not want to collaborate? (Laughter)

MR. COOK-DEEGAN: If I could maybe sort of short circuit this whole -- This was the topic of our last workshop, and we will have sections on this in the report. I think that it's not really the topic today. I don't think it would be productive to go too much further on it.

MS. DONIS-KELLER: Let me just say the last word.

CHAIRMAN BERG: No, I really wasn't trying to put you on the spot. I think, appropo to the question is how are we going to use this information? We'll have a map. The map consists of probves, and the question is, if it's going to be used widely and -- It's not just your collaborators, but it seems to me it has to be available to people who want to explore the problem from whatever point of view they have.

MS. DONIS-KELLER: I couldn't agree with you more. Of course. There are many ways to accomplish that goal, and Collaborative is very interested in advancing the course of basic research. We need to find a mechanism so that the probes can be widely distributed for any purpose whatsoever, provided that Collaborative does not lose what it has invested.

That seems only fair to me. We also -- The other points that I would like to make is that the information is being given to -- collaboration, and it's being published at a very rapid rate. So the information on where the markers is there, and that information has a lot of value to other people who are using the markers themselves in their various laboratories. They can know exactly where their markers map with respect to other genetic markers. So they can home in on regions that they're interested in without even having a physical probe. They don't have to duplicate the work, because the work has been done on this set of reference families.

Okay. So that information is a very powerful resource in and of itself, and we are already making that available through the collaboration. So the issue of the physical probes themselves and their distribution, I think, is a very solvable one.

CHAIRMAN BERG: Okay. I'm going to terminate this part of the discussion and ask Bob to summarize what he thinks the numbers are and whether he's got enough answers to the three questions, that is now changing the 10 to a 5, and whether you think you've got enough information to cost out the infrastructure.

MR. COOK-DEEGAN: I think we're going to have to still figure what the facilities costs would be and what the long term commitment to repositories and data bases would be; but aside from that, I think we've heard a very solid figure for a 5 centimorgan, somewhere on the order of 11 million, and a lot of the -- that probably somewhere between five to ten times that size for a 1 centimorgan, with a question mark about whether that's a legitimate goal, and almost de-linking it from the rest of our agenda after you hit the 5 centimorgan level because of questions about how much it would be useful in ordering a physical map.

The other thing is, I guess the other point is the main purpose of doing the RFLP map is it's a gateway to making the tools that we're going to talk about for the rest of the day useful. That is, to link an unknown disease locus and make it possible to do the -- to take advantage of the physical maps and the other technologies we've been talking about.

MR. BURKS: Could I ask just one quick question? I didn't hear anything about experimental resources as a cost.

MR. GILBERT: You mean consumables?

MR. BURKS: Yes. Enzymes and --

MS. DONIS-KELLER: I still say 100k a year.

CHAIRMAN BERG: That's in the 100,000 a year.

Okay, I think we'll move on and move on to --

MR. BURKS: Could I just say one thing, Paul. The real question is, do we want any other -- to encourage any additional groups that might make significant contributions in this matter? Then you get into the question of a physical plant. If you want another group of twenty-five, you probably have to create the space and the instrumentation. Or do we want to just have these two groups that are already doing it finish it up for $11 million?

I think you probably want to build in some expansion in the number of people that are doing this.

MR. WATSON: Could I ask, you mean -- The difficult question is - The Hughes effort and the Colleaborative effort working hand in hand, or --?

MR. CAHILL: It's difficult to -- because we've had discussions now for two or three years and, obviously, one is a commercial venture, the other is a totally public venture. For example, of the --

MS. DONIS-KELLER: It's a private institute.

MR. CAHILL: It's a private institute but by nature tax exempt. Everything we do is in public domain equally -- I've got my counsel sitting over here. Correct me if I -- (laughter). But to get back to the topic, the CEPH families, of the forty CEPH families, 27 have been collected by Hughes and put in that domain, just to give you an idea; and they're for the public use at large. Two are Amish. One is Nancy Wexler's Huntington's kindred, and the other ten are blue-blooded Frenchmen, or I should say "tricolour" blooded Frenchmen. And that's the background.

So that we're trying to get it out into the public domain, but when there's a very highly informative probe that Collaborative developes, it's in their personal thing to hold onto that until they see what diseases are correlated, and hold onto it as long as they can, which is proper. It's a business venture.

So that the probes that are let out, if I'm correct, are let out when there's a specific disease in mind, a hot area, in a collaborative venture between Collaborative and the investigator; but they're not let out for the use of the rest of the people in the world to establish them at the moment, if I'm correct.

MS. DONIS-KELLER: Just to give a few examples, we have let out some of these very highly polymorphic probes for pure basic research. It's not only the ventures on the disorders that affect the largest number of people. For example, the XY probe has a PIC of .98. We've made that widely available to any researcher for research purposes.

In fact, we're very interested in promoting basic research, and our probes are being made available; but you're right, there is this problem of the private monies that finance that resource.

CHAIRMAN BERG: Okay. I'm going to have to go on, but I want to --

MR. CAHILL: Paul, one -- If the probes of Collaborative were to be tied in with the ones that are already present in the Utah group, how close would it be today having the 5 centimorgan map today? In other words, in many areas it's already done.

MS. DONIS-KELLER: Well, since the Hughes Institute hasn't published -- If they have a map of several hundred markers, it's not out in the scientific literature at the moment. From what I understand from the work that's being done, I think we would be reasonably close to having a 10 centimorgan map today.

I would also just like to say, though, we have made the information available through the CEPH collaboration, and the information on our genetic map is available to the entire academic community. We do give out probes, highly polymorphic probes, for people interested in basic research issues. We don't share them with other companies so that they can get into --

CHAIRMAN BERG: Okay. I want to touch one more item just in that. How good do people believe this estimate is 100 k per year per person? Is that a fair one, because a lot of the multiplication of all those numbers depend on that estimate. If it's off by a factor of two, then clearly --

MR. CAHILL: Well, I think they're a good fidelity, because what Helen presented is close to what our budget for doing parallel kind of work, number of people, in progress. I guess the Salt Lake probe is about equivalent in number to what you have said. The fact that the numbers come out the same, I think that's a pretty reasonable guess.

CHAIRMAN BERG: And that's not just specifically for this work, I assume -- I mean, what it costs to fund work at --

MR. PEARSON: duPont charges -- They're significantly higher than 100 k per year, but, of course, all of this then is going to depend on the quality of person that is being applied to the problem. And, clearly, that's -- You can go with less skilled people later in the program when you know exactly what you're doing and it becomes a crank turning problem.

MR. GILBERT: I think we should be careful about that. I used to figure 100 k per year was adequate for everybody that's in the place, including the janitors. As a way of getting a cost, it includes everybody that's actually involved in it. If you use that figure per PhD scientists in a group project, it's a very different figure. So I'm not sure what you're thinking of.

MR. PEARSON: Yeah. I agree with you totally.

CHAIRMAN BERG: So it's going to be more, is what you're saying.

MR. GILBERT: No, no. Counting how many bodies. Get an idea of how many bodies you got to put in the building to do something, including secretaries and the janitors and then multiply that by $100,000.

CHAIRMAN BERG: Is the janitor included in your manyears? No.

MS. DONIS-KELLER: This is just scientific people on the question, and it's a minimum estimate for overhead. Everyone calculates a little bit differently, and it depends on how many administrators you have or vice presidents in universities. I mean, as far as what you calculate the overhead for MIT might be versus some other --

CHAIRMAN BERG: That's why I say, the number of 10 million could easily be 25 or 20. I mean, if you're talking about scientific people which is the number --

MR. CARRANO: I think you're really talking between 100 to 150 k per person. I mean, including all the supples expenses and variable overhead.

MR. COOK-DEEGAN: We'll change that 10 to fifteen or something like that.

CHAIRMAN BERG: And then maybe the factor to go up to the 1 centimorgan might begin to use, let's say, lower scales to just do repetitive jobs that maybe it isn't quite as big an increment. So -- But I mean, the lower estimate for the 5 centimorgan then is maybe ten to $15 million. Okay. I just wanted to get that number so that we can -- Of course, we're going to be using that number over and over almost at every stage.

MR. McKENNEY: Perhaps it's more important to record the number of manyears and come to agreement on that.

CHAIRMAN BERG: Well, somebody wants a number with a dollar sign in front of it.

MR. McKENNEY: Well, at the Bureau of Standards they use a value of 2.5. So you take the salary and the fringers and multiply by 2.5. That includes all the overhead, the maintenance but none of the mortgage or rent. I think that that's considered cheap in the industry. I don't know what it is at duPont, but I would say it's probably in excess of 2.5.

MR. PEARSON: Yes. But I would like to come back to Wally's point so that we're careful in these estimates to make sure that we're not talking about preexisting infrastructure and an incremental cost always, because Congress can be too readily misled. The other issue to be careful about is overhead and labeling things as dollars in overhead or percentage cost.

It was a great shock to me, going into industry, to find that the calculation that was used for overhead assumed as 100 percent of the cost the direct cost, and to find that in some cases, in fact, the issue was reversed, namely, what is considered 100 percent of the cost is both direct and indirect or overhead plus direct. And you can end up with differential estimates of overhead that on a percentage basis look like they differ by a factor of two, but in reality the absolute number of bucks involved in getting the work done is the same.

So we have to be careful of that kind of semantics.

CHAIRMAN BERG: Okay. Given all those cautionary comments, we'll go on to the next thing. That is to define the physical map. There, I think we -- In a previous discussion there seemed to be some difference of opinion on what we meant by a physical map. It is in fact to obtain an overlapping set of cosmids which contain all of the DNA sequences of the human genome, in this particular case, not just knowing where there are some restriction sites. It is in fact to obtain a library of all of the DNA in varying lengths, fragments, available so that you can line them up.

I'm not sure I know of anybody here who's actually been engaged -- Tony? I'm going to let you talk to that issue first.

MR. CARRANO: I can speak specifically on the production of cosmid libraries using flow sorted materials. So we're talking then about chromosome specific ordering and the use of cosmid clones from individual chromosomes sorted by flow cytometry.

Let me address that issue, but before I do I want to -- An issue that impinges on what I'm about to say is that little phrase by Helen, and that is in part of her estimates for the 750 RFLPs she included the construction of cosmid genomid and single chromosome libraries. Was that really meant as part of that estimate?

MS. DONIS-KELLER: The phage maps and cosmid maps are very easy to construct, as people in the laboratory can make those maps. The single chromosome libraries I was thinking of are the libraries when you're provided with a DNA that's already been flow sorted. It does not take into account making a very large chromosome library. So that would be in that. What you're talking about --

MR. CARRANO: Is a separate issue.

MS. DONIS-KELLER: -- is a very separate issue.

MR. CARRANO: Our estimates would be quite different, I think, in that regard.

Let me address the production of the chromosome specific libraries using cosmids, and this is an issue that both Livermore and Los Alamos are currently engaged in. As you know, we have completed all of the human chromosomes and complete digest and Charon 21A vectors, and they are currently in the ATCC repository and now being distributed and heavily used.

We're now turning toward the production of cosmid libraries, and my assumptions in this is that there is no further development work needed. The one obstacle that we had in the creation of large insert libraries has been in flow sorting chromosomes. That problem has recently been resolved with the modification of the chromosome isolation buffer that is being used.

So in fact two large insert libraries have been created at Los Alamos, one on X chromosome and one on chromosome 16 and charon vectors, charon 35 and charbn 40. And both laboratories are currently in the process of making cosmid libraries.

The present plan is that each laboratory, Livermore and Los Alamos, would take half the human genome. The exact chromosomes each laboratory would get are still to be decided, but the initial sets have been chosen.

The project in terms of the flow sorting can be divided into several steps with different estimates in terms of person power or personyears for each. It turns out that a major effort that we discovered in the creation of the charon 21A library has been in the cell culturing and the chromosome isolation.

Cell culturing has really become an issue. Maintaining the cell cultures -- Most of the work for the small chromosomes has relied upon hybrids for flow sorting. Obtaining those hybrids or constructing those hybrids has been a real issue, and then maintaining them in culture without loss of the chromosome also has been a fastidious -- requires a lot of care.

Our estimates for the cosmid libraries now, assuming that we have most of the hybrids that we want to use or the cell lines for the large chromosomes from the human, is approximately ten personyears for the cell culture and chromosome isolation work.

The flow instrumentation, which includes both the maintenance and the actual physical sorting of the chromosomes, is about twelve personyears. For the DNA extraction and cloning, which are not trivial because we are talking about cloning tens of nonograms worth of DNA which are available from flow sorting -- The extraction and the cloning efforts are again about a twelve personyear effort.

In the past for the charon 21A libraries, because of funding constraints, we had made a decision, and also at the request of our Advisory Boards, that we do not do any major characterization on those libraries; and as a consequence, we have not. We have just done very minimal characterization, which is basically looking at the frequency of nonrecombinants or the percent of non-human contaminatient present in the libraries that were created.

We had anticipated that the scientific community would give us feedback. Unfortunately, that feedback has not been easy to obtain, despite the fact that many -- several hundred libraries have been put out to the scientific community.

I'm going to assume again that we will do minimal characterization on the cosmid libraries, and then I'll give another estimate for additional characterization. But if we assume a minimal characterization for the cosmid libraries, that is again looking at the frequency of nonrecombinants, the frequency of rodent contamination sorting from hybrids, and just minimal characterization in terms of the chromosome specificities, just mapping a few probes from that liabrary against a panel of hybrids that will tell us whether or not we've got ten probes located on the chromosome of interest, we're talking about a ten personyear effort.

The quality control that goes into the effort is both in the area of cytogenetics in terms of what material goes into flow sorting and what material comes out of the flow sorting, and also on a mathematical analysis based upon the histograms obtained on the flow system. And this is about a six personyear effort.

If you add all this up, it comes out ot about fifty personyears to construct cosmid libraries and each of the human chromosomes. Then you can generate the number that you desire to use. I chose to use 150 k per personyear. That comes out to about $7.5 million.

There is also some associated equipment that goes with that. Lasers, of course, are a major component to the flow systems; and they have to be replaced periodically. I would estimate that approximately $1 million is needed at least for that component of the equipment for the laser operation.

So we're talking about about $8.5 million, and I think we have leeway of at least fifty percent on that. So we're talking somewhere between $8.5 million to about $13 million.

CHAIRMAN BERG: The leeway is all up.

MR. CARRANO: Yes.

MR. McKENNEY: Fifty personyears per chromosome or for the whole --?

MR. CARRANO: For the whole job.

CHAIRMAN BERG: Now that's just to obtain the libraries. That's not ordering them.

MR. CARRANO: That's correct.

CHAIRMAN BERG: That's just talking about libraries.

MR. CARRANO: That's correct. Now our estimates -- Our plans right now -- Again, this includes no development. We've done most of the developmental work initially on the charon 21A libraries, and we think we can proceed with minimal amount of development except for cosmid characterization.

Our goal right now is to complete the entire human genome and cosmids within the next three years. That's the plan.

CHAIRMAN BERG: Next three years?

MR. CARRANO: Within the next three years, flow sorted.

MR. HOOD: Tony, as you know, I was on your Advisory Board along in earlier stages, and have not participated since. But I've heard frequent criticisms or charges which essentially says that the libraries are very frequently not what they are supposed to be, that they're contaminated -- I mean, there's been a lot of things. I am not surprised if that was true of the earlier stages, because, frankly, an effort of that magnitude and complexity with a group of people who weren't totally involved and inexperienced in that. But what is the general feeling now? I mean, are you very confident and are most people happy with the quality of the libraries that you are making?

MR. CARRANO: I would say that there has been some expressions of -- not necessarily dissatisfaction. There have been some expressions that the libraries, at least two of them, have been contaminated heavily with a rodent background, particularly hamster; because some of those libraries were sorted from hamster-human hybrids.

However, I presume you've also seen a rebuttal to those issues by a prominent group of scientists, both in Britain and in Canada, who argue very strongly that that's nonsense. Those libraries are still very valuable. Even if you're talking about 50 percent contamination, you've saved yourself an effort of over tenfold in terms of screening the libraries or clones of interest.

That represents not human error. That represents state of the art technology. Some of the hybrids from which those chromosomes are sorted, the human chromosomes are not in favorable positions to separate them purely from the hamster background. So that really is state of the art technology. They're not human error mistakes that are made in the process.

We try -- We tried to obtain newer hybrids that allow more favorable separation of those human chromosomes from the hamster background, and those libraries that initially we were concerned about, chromosome 16, in particular, have been redone, and the quality of the new libraries is much better.

When we present these libraries out to the ATCC, there are spec sheets that go with them that explain in fact what our estimation of rodent contamination is. The user has the option of not using that or using it, as they so please. But it really is state of the art. Nothing much we can do about that.

MR. HOOD: Tony, how would you see the rodent contamination which would presumably still be present in many of the libraries, in fact, in this whole project of physical alignment of the cosmid clones if you're going to have to rely on repetitive sequences as probes to say that this fragment DNA is from humans and then -- Do you run into problems with regions that don't have sequences?

MR. CARRANO: Yes.

MR. HOOD: I'm just curious about kind of how you see tactically this impinging on --

MR. CARRANO: It's a very good question, Lee. I'd like to see what John's answer is on this as well, but in the libraries that we're currently making our goal is to obtain better than ninety percent or ninety-five percent purity, that is, human material, human material of interest. So the real question becomes do you want to then take a library and screen it with human repeats to identify the human clones and pull those out, or do you want to ignore that in the ordering process and let the eventual ordering and sequencing screen out the rodent sequences for you.

It's not clear to me what's going to be most cost efficient. So I think we're going to have to work that through as we go along. I think we have the option of doing both.

CHAIRMAN BERG: According to Charlie Cantor's comment, they have obtained megabase DNA by their -- which do not contain repetitive sequences. There seem to be fairly long regions which don't have the kind of repetitive sequence that we use as being diagnostic with human. So, clearly, that criterion alone might not be sufficient. John, you've had a large amount of experience in this area.

MR. SULSTON: Yes. I've sort of caught unprepared. I was -- Do you want me to very quickly say what our experience is, and we'll all multiply -- (laughter)

CHAIRMAN BERG: Tony has said whta he thinks it will take to get us to a library stage. And you've obviously gone beyond that in learning how to link up large segments.

MR. SULSTON: Okay. Well, I'll approach it that way. Maybe I should just speak to this last point. I mean, I think, within reason it doesn't actually matter if you have five percent of rubbish in there, because it will come out in the wash, provided your linking is working anyway. You'll end up with some stuff that you can't place, and then eventually find out it's hamster. I think we're probably in that position, you see. The caenorphabditis nematodes that we use are grown on coli, and I fully expect that a few of our contacts may turn out to be coli, and that in due course will screen out. But for the moment, there's no point in doing so. It will take care of itself.

So very briefly then, our experience has been with this nematode, caenorphabditis which has an 80 megabase genome, and that's distributed over six chromosomes of roughly equal size. And because they're small and because they're roughly equal in size, we never contemplated trying to sort them or separate them in any way.

However, we've got a method that seems to be perfectly satisfactory. Of course, conveniently, we are talking order of magnitude about the single human chromosome here. So the technique is going to be comparable if one wants to repeat them.

We found, working with -- I've aimed for basically a shoestring mark with the primary emphasis on archiving and providing genetic communication for the largest number of laboratories who are working on genetic and now molecular characterization of regions of the genome which were in development. So we got started fairly slowly with the map, and only now, with current technology, are we seriously thinking about -- that we may finish it sometime.

In fact -- One of the most irritating things about the present situation is that people never ask how is the map going. They say, what are you going to do next. Anyway, the -- So what we've been doing is to take random cosmids from this 80 megabases. We've been calling it a stereotyped restriction and separation of radioactive fragments. Then we have a data base and computerized digitizer and so forth.

The most useful number I can give you on costs, I think, is what it would cost if the lab burned down while I'm away and we had to start again tomorrow, how long it would take us to get back to where we are. And I think it's two personyears.

It turns out that, rather appallingly, that when I add up the costs, they're 90 percent salary and, really, the running costs are pretty small, well, perhaps 80 percent salary, to be fair, which, of course, reflects the fact that we're seriously underautomated. So one would not scale up precisely what we're doing.

However, if one takes then two personyears to get back to where we are now -- and where we are now is that we have cloned, I think, about 90 percent of the genome, maybe 95 percent. That's based on the frequency with which other people's clones selected either as genes or as RFLPs of one kind or another -- the frequency of which those clones fall onto our map into known regions, and that's 90-95 percent.

Now the remaining 5 - 10 percent is presenting more of a problem, and this is a very crucial thing to consider, which is that cosmid banks, in our experience, are by no means random. Now this may be overcome by one means or another, but of all the means that have been proposed so far, none, I think, have been totally successful. The most useful development since we began is Peter Little's development of the cloning vectors which are based on the Lambda origin of replication rather than the -- That goes only a little way towards the problem, I think; and we really think that we are -- we have a lower instance of some clones.

That doesn't mean to say that there's anything which is necessarily unclonable. There may be; there may be not. But at least there are things which are sufficiently underrepresented that the random approach runs out of steam. So what we were expecting to do at this point where we are now, to get the linkage together. we jave abpit 700 contigs at this point. So in round numbers we have an average contig size of 100 kilobases.

The maximum contig size at the moment is one megabase. So that gives you a feeling for where we are.

Now to link these together, we thought we would have to do probing either into cosmids or, if we really run out of steam with that, into lender libraries. Now, of course, this year the goalposts have moved very substantially with the development from Maynard Olson's lab of cloning in yeast; and we're at the moment getting our hands dirty with that. We're beginning to make progress, and I have every expectation that will allow us to finish in a reasonable time; because we expect, as has been pointed out by Maynard himself and others, that not only are the X big but also we expect that they will provide more random cloning of eukaryote sequences, because the DNA in the yeast is single copy. It's protected with histones. It's growing in, as it were, its natural environment. It's a completely different situation from trying to clone relative naked eukaryote DNA in coli which, I think, in some cases is really leading to problems.

Right. I think that summarizes -- Well, I suppose, and then -- I mean, very optimistically. I mean, I might expect that if the X goes well, we should finish in a year. So it would be two more personyears. But I mean, that's still projection, and I can't say --

CHAIRMAN BERG: How do you know you're missing sequences in your library? I mean, how do you make an estimate of what you're missing or how much you're missing?

MR. SULSTON: Well, to reiterate, we're starting the ground on cloning. We're building up contigs, which I haven't defined, but I think everybody understands they're interlocking clone groups, just by matching. Now, of course, you don't know. You have no idea what tracks are left in between. We have two ways. One is approximate. Many years ago I made an estimate of the nematode genome size, and that's where this 80 megabase figure actually comes from. And that may be off, of course, by 10, 20 megabases.

So that, anyway, gives us a ballpark figure. That's all we can say. Then this other thing of taking clones which other people have selected and putting them on our map, we find that 90 percent or so of these incoming clones do, in fact, fall on the mapped regions.

So that suggests that there may be 10 percent still to find. But we don't know in any other way how big the gaps are until the map is finished. And that's why the yeast cloning is exciting, because arithmetically it will allow us to bridge across gaps of larger size, whereas before we were having to anticipate in awkward areas really doing hard, grinding Lambda wards to get through. But we don't know, Paul, I'm afraid, and it's just these soft numbers at this point.

CHAIRMAN BERG: And another technical question is: Given the fact that, with the recent advances in the resolution in pulse field electrophoresis of the various groups, why isn't hybridization to these kinds of gels a way of linking -- you don't need to clone them.

MR. SULSTON: Yes. I'm sorry. I gave you the Lanbda. You're absolutely right. The other thing which we sort of thinking about, the so-called matro restriction marking. It's not immediate and direct. If you take coli and you cut it with not and run a pulse field gel, then, of course, you have narcistic fragment. If you take even something nematode size, then you have a smear of 200 fragments.

So you cannot with a single not cut and pulse field -- you cannot immediately say whether or not two things are linked just because they go to the same place. So you would have as a minimum to do two orthogonal things. I mean, say, SFI, for example. And then you would still be left with a stochastic answer. Now much more powerful if we can get hold of the X, because then you have the thing in hand. You know it's unique, sure it's unique. So that's why it's exciting.

That's why it will speed up. You're quite right. I'm sorry. I forgot to mention the micro restriction.

MR. HOOD: There have been two general strategies discussed for physical mapping of the human genome. One would be what Tony has said, separating the chromosomes and starting from those; and the second would be to take the entire human genome and construct vectors. What are your thoughts on those two alternatives approaches, John?

MR. SULSTON: I would be completely pragmatic. I think that you have to consider the quality of the banks, and you have to work out the -- which is the point you addressed -- quality of the cosmid banks coming from the chromosomes and consider whether that's going to save you much in time. But there's one point I should make, which is that I would think that with twice the amount of analytical effort that we use, that one should in principle be able to place a clone anywhere in the human genome; because just to think about it statistically in a very crude way, we are placing a clone using a single lane analysis. We're placing a clone in 80 megabases of DNA which would be covered by

2500 cosmids approximately.

Now the human genome is forty times as big, and so statistically I would guess that two lanes passed through a computerized coincidence detector, as it were, would be sufficient to uniquely place that somewhere in the human genome.

So that's the thing you've got to bear in mind. So that's not -- If that's true, and of course, that's based on a number of assumptions -- But if that's correct, then it's less than twice the effort, because you've already made the DNA. You simply do two analyses on each one. Make it three, if you like; and you still aren't putting out a lot more effort, especially as that's the part, of course, that's most susceptible to automation.

You've got the clone. You made the DNA. Now you just do biochemistry of some sort of machine, other sorts of things that you've been talking about. This is all into the automation.

So I do think it's very important to bear this in mind. But on the other hand, the cost -- I mean, I said we made a shoestring map. I also say we're making a very pragmatic map. I've been very conscious of the fact that we're feeling our way, and I've always been very loathe to get into very accurate estimates. I always wanted to keep the computer a little bit to the background. It just does the easy stuff. We don't have very fancy computer programs. We use our eyes on the films, and so forth; because we don't know where we're going here.

We don't know what sort of strange redundancies are going to turn up in the whole genome. And so it may be, you know, just a statistical argument like that just won't work, in which case the chromosome sorting could be terribly important.

MR. CARRANO: John, in light of this very question, repeated DNA in the nematode versus what we expect to see in the human, how much will that --?

MR. SULSTON: Okay. The equivalent of a middle repositive from DNA in the nematode is in fact a broad spectrum repeat. There isn't a two step -- There's a great smear running out from twofold 'way out to a thousandfold, and the total repositive DNA is around 15 percent of the genome.

MR. CARRANO: Fifteen?

MR. SULSTON: Fifteen, yes. On top of that, there's a -- there's foldback DNA, about three percent.

Now that -- In the nematode we have zero problems from that. In other words, the -- in the length of the cosmid the combination of inaccuracy of the repeats and small size and the dispersion of the repeats is such that we simply do not have, for the most part, ambiguities from that source.

There are one or two places which have been set aside and are interesting where we've clearly got tandem repeats, and it's hard to sort out the map in those places, and we just leave them with a note on the computer at the moment. But even those, I don't think, are providing -- are causing problems from one place to another.

We are also -- and we're getting increasing is the point now where we're really checking objectively, because we're joining together things more and more that Don Albertson has already placed by in situ hybridization. So, I mean, quite apart from my own feelings of it, I can say it does actually work.

Now the thing is, what I'm afraid I'm totally ignorant about is how to extrapolate this to the human case. I would think it might not be terribly difficult, and that again the inaccuracy and the dispersion of the repeats is such that you'll have precision. You'll have unique placing of the restriction enzyme level.

What bothers me more, I think, is a point that nobody really knows about, is the question of heterozygocity of the source.

CHAIRMAN BERG: You addressed the issue of clonability.

MR. SULSTON: Yes.

CHAIRMAN BERG: But that's one issue. You may not be able to get certain sequences of DNA. You may. But on the other hand, there's certainly enough instances where rearrangements that occur on cloning have been noted, and clearly you will find that as being a sequence that you cloned, but it's not really the native sequence. And it seems to me, that's not a trivial concern.

MR. SULSTON: I don't think it's trivial at all, and I don't know whether -- I don't know how well cosmid cloning technology can go in the future. I would be sort of surprised if it eventually pays out. If the yeast clones are good, then we may well use those for cosmid size ranges as well, because we know we've got problems with some cosmids.

CHAIRMAN BERG: But even in yeast it's not clear to me that you have -- Maybe there are strains of yeast in which the frequencies in combination is exceedingly low and you don't have to worry about it; but if you put long pieces in with direct repeats, you can easily, it seems to me, get rearrangements which you would not right off know had occurred.

MR. SULSTON: No, no, no. I'm sorry. I say, if the yeast goes well. By that, I mean, if that doesn't prove to be too severe. But I mean, I want to be very honest and say that the cosmids are not totally satisfactory, by any means; and we are having problems with labs complaining because they've got stuff with deletions and so forth.

Now what we do is to -- You know, we've always known that this might happen. So we've been very careful to keep our primary things frozen. It just means, basically, extra mailing effort, because -- When people aren't very skilled at growing them up straight away, you know -- Of course, these things show up. But it means that I would not, in the nematode at least, advise anybody to do sequencing from those cosmids at the moment.

Now it may be that we'll learn more about how to handle it. It may be that in the future better vectors will arise, which would use those. We don't know, for example -- It's a real problem. I mean, I don't know -- I don't think that I would say that our cosmid bank is a good place to start sequencing the nematode right now.

CHAIRMAN BERG: An interesting region which might be used as a model is the globin region, the Beta globin region, which is some 100 kb in length with lots of allele repeats and different orientations, and trying to see if one could clone that region and whether remains -- Since the whole sequence of that region is now known, it's an interesting model to ask whether you would clone the length of that size knowing some of the problems that are likely to be encountered in the rest of the genome.

MR. SULSTON: I don't know whether it's going to be related to particular things like that. For example, we've not had a lot of trouble, at least with most of our banks, with the ribosomal tandem repeat. You see, that's about three -- And it doesn't normally delete, you see. So it doesn't have to happen. I don't know why we're getting -- And it -- I mean, it's not across the board. I mean, the point is that it has not provided us with any great problems. The map is going well in terms of the matching.

What happens is that, if somebody grows up a clone, having left it lying around for a while, then sometimes they will find rearrangement. Okay?

MR. HOOD: I think a critical issue this speaks to,if you're going to start sequencing the human genome, it's really critical that you have normal DNA from the individual available, so that when you got into the mapping and into the sequencing, you can go back to that individual and you can pull DNA out that you know is normal, has not been in tissue culture for five years and these kinds of things. So you need that as an ultimate check for these kinds of questions.

So getting DNA from --

CHAIRMAN BERG: We need to freeze somebody?

MR. HOOD: -- these individuals who have long since been dead is not the way to go.

MR. McKENNEY: John, would you comment on the use of Lambda as opposed to cosmids? It's clear that the chromosomes are going into cosmids now.

MR. CARRANO: The chromosomes as opposed to cosmids here -- The chromosomes are actually being made in two vectors. The same material that's being sorted is actually being put into charon 40 as well as the cosmid vector. So both libraries will be available, both Lambda and cosmids.

MR. SULSTON: I think you need -- I mean, I think that we've gone the right way for current technology in making our map; but at the moment in the nematode it's cost effective, I think, for a lab that wants to sequence a particular region to go and put out genetic lambdas using the cosmids as probes. And that's very quick and easy for them to do, you see. It's getting the map, getting the linkage, which is crucial; and for that you want the big clones. We certainly want the X. We may continue to want the cosmids for quite a while. But what now -- The thing at which I'm lost and the thing that should be discussed is whether one should concentrate going the same way with human or not, and say, well, okay, we made the map and now we pull clones as we need them. I mean, maybe -- In fact, it may be even in the human it's not that much work in the whole spectrum of things to do that, so that we use the big things to do the linkage and then go and get primary genomic lambdas. And that, to me, is a very safe way to go. So you end up sequencing lambda clones. Does that make sense to you, Lee, to do that?

MR. HOOD: That certainly is a possibility.

CHAIRMAN BERG: Trying to focus back on this cost estimate, let me just see if I can return. Tony, you talked about making an entire cosmid library for each chromosome. Is that correct?

MR. CARRANO: That's correct.

CHAIRMAN BERG: Okay. And you gave an estimate of the cost and, if I remember correctly, we were talking about of the order of $10 million.

MR. CARRANO: Yes.

CHAIRMAN BERG: Is there a plan in your shop to go on and try to order them, and can you give us any kind of estimate that you would arrive at of what it will take to order those cosmid clones, having heard John or knowing of John's experience and other people's experience? What do you -- Can you help us on that?

MR. CARRANO: We are interested in chromosome 19, and our interest is not initially in doing the entire human genome, but for genetic issues related to chromosome 19. In ordering chromosome 19 you're going to order a set of cosmid clones for that chromosome.

We have been underway in this project for a little over a year right now, and most of that has been developmental work; and within the next two months, we're actually beginning the production work. But it also relies on automation. It relies on the adaptability of the technology developed by Lee Hood and Smith to basically do all the gel analysis and reading automatically.

CHAIRMAN BERG: A sequence.

MR. CARRANO: It's basically applying the DNA sequencing machine to do the mapping. So that puts an element of timesaving in it.

Before I go up to the estimate of chromosome 19, let me come through some other estimates here that I made some calculations on for a while. Some of us were in Japan at the meeting and heard Dr. Kahara speak about establishing the complete physical map of E. coli in lambda, and he's got essentially the whole thing ordered with the exception of three gaps which are basically uncloned regions.

MR. McKENNEY: I thought it was six.

MR. CARRANO: Six.

CHAIRMAN BERG: Six unclonable regions? Six gaps? Is that what you're saying?

MR. CARRANO: Yeah, six gaps. Unclonable. In his comments to me, and I think several of us were there, was that it was really a six manmonth effort to do the entire E. coli genome. He did it himself, basically, in six months. And if you just extrapolate that to the human, it comes out to about 300 man years of effort, just scaling up linearly; and that may not be the correct scale.

Our estimates with the automation for chromosome 19 would be about two man years worth of effort. So if you scale up to the human genome, approximately 100 man years of effort if it's a linear scaling.

CHAIRMAN BERG: Where does 19 sit in the size?

MR. CARRANO: Well, it's one percent of the human genome.

CHAIRMAN BERG: One percent. So it's small.

MR. CARRANO: Right.

MR. WATSON: What's the technique you're using?

MR. CARRANO: The technique, basically, for gel reading and data analysis, is based upon the principle that one can generate much more information on cosmid overlap by using high density restriction fragments. One approach has been Dr. Kahara's approach, to use a series of six cutters. He used basically 8 six cutters to generate -- and partial digest of the system to generate overlapping set of lambda fragments.

Our approach is to use a combination of four cutters which would generate a high density restriction map of, say, from the ten to 400 base pair would generate approximately 50 to 100 fragments for each four cutter digest. Those can be actually coupled -- Each of those fragments could be coupled with flourochrome and can be run basically on the sequencing apparatus that Lee Hood and Lloyd Smith and Mike Hunkapiller developed and can be analyzed directly on that system running three cosmids per lane with an internal standard in each lane.

That, in fact -- If you compute that through a chromosome 19 for a 5X library, it results in 110 man days worth of effort assuming no mistakes are made along the way. The data are output directly to any work station for data analysis. So one can essentially -- it's very simple to do that.

So just looking at straight machine time without the biology that goes behind preparing all that DNA, you're talking about 110 machine days.

MR. HOOD: And the critical point, Jim, is you have one base pair resolution up to three or four hundred. So it means you can really look individually at these 400 fragments or so that you get from a typical cosmid flow. So it's a real fingerprint. And then the four color single lane analysis. But you have an internal standard that you can compare every single clone you look at back against that internal standard, so it in turn can be prepared against every clone you analyze. So it is -- It really changes how you map in a very powerful sort of way.

CHAIRMAN BERG: So to get back to some numbers here, your estimate was -- The Japanese effort with intensive labor, 300 man years. With the improvement that you describe, it goes down to 100 man years, and that's an extrapolation from what you estimate for a single chromosome 19.

MR. WATSON: Right.

CHAIRMAN BERG: Anybody got any better fixes? Yeah?

MR. McKENNEY: Could I just ask, are you starting with DNA, or does that -- Are you starting with purified clones for your analysis using your automated system?

MR. CARRANO: We're starting with cosmid clones developed in a modified lorus vector.

MR. COOK-DEEGAN: But one step back. I think he's probably asking the question I asked you a couple of days ago, which is the source of DNA.

MR. CARRANO: Okay. We're dealing with DNA from monochromosomal hybrids. Okay. However, we have the donor available to us. We have established cell line. We also have the donor for freely willing to donate further samples if we so need it.

MR. McKENNEY: My question is, you obviously have some DNA. You digest it. You tag it with your flourochromes. Does your estimate include the preparation of that?

MR. CARRANO: No, no, no. The 100 man years includes that. The 110 machine days does not include that.

MR. McKENNEY: Can I say something else? The issue about the choice of DNA and having the original source available -- It's not clear to me that that's an absolute requirement, because it's not clear to me that we really are going to do the sequence on the DNA that we initially order. I mean, there are schemes available for taking the DNA that is ordered and going back to any genome that we want and pulling out a similar piece of

DNA to do the sequence from that individual. And that may be a strategy that's worth considering as well.

MR. CARRANO: I think for genetic purposes we won't want to do that anyway. We're not going to want the DNA from one person all the time. We want to do some genetic studies.

CHAIRMAN BERG: The eight to ten estimate that you gave us to get the clone library pales into insignificance when you start talking about going on to the ordering. So, in other words, it's roughly a round number. The lower estimate is 100 million --

MR. GILBERT: That's not same figure, Paul.

CHAIRMAN BERG: Pardon?

MR. GILBERT: It's 100 man years.

CHAIRMAN BERG: A hundred -- No, no. I thought the 100 man years was on top of the 50 man years that he --

MR. GILBERT: Yes.

CHAIRMAN BERG: Right. So it's 150 man years.

MR. CARRANO: Talking about the production of the cosmid libraries plus ordering, it would be 150 man years.

CHAIRMAN BERG: With the technology improvement that you mentioned. Anybody else got any --?

MR. PEARSON: Just one comment on the technology improvement side. I think there is reason to believe that we can do much better than that, just to extend the current applications. Therefore, that will reduce this estimate of 100 person years to do the genome.

MR. CARRANO: I would concur with that. I think that certainly can be reduced with improved technology.

CHAIRMAN BERG: Now you gave us -- Okay. So you're talking about --

MR. COOK-DEEGAN: On the order of $10 million again. Right? What was your multiplying factor? Oh, excuse me. Double that. Sorry.

CHAIRMAN BERG: That's to get a completely -- let's say, as nearly completely as one can get overlapping libraries of human DNA segments on all the chromosomes.

MR. CARRANO: That does not include any characterization work. I mean, that's just laying them out perhaps 80 percent complete. It doesn't -- It's not including going after those final clones to close all those gaps, which is --

MR. GILBERT: Wait. Your two years of effort-- What is it?

MR. CARRANO: We're talking about 80 percent. For that, we can probably achieve 80 percent closure, which is -- I think --

MR. GILBERT: Eight percent is nowhere near closure.

MR. CARRANO: I can't give you an exact answer. Our modeling of this effort indicates that we should be able to achieve 80 to 90 percent closure.

MR. GILBERT: And how many covers? How many times are you covering?

MR. CARRANO: At that level you're covering about ten times. You're going through it tenfold.

MR. GILBERT: And you feel it will only be eighty percent closure?

MR. CARRANO: Right. The modeling experiments pinpoint it. We've modeled John's nematode experiment, and we come smack on what they have come up with. We modeled Maynard Olson's experiment, and we've come up just about smack on what he's found in terms of modeling effort.

MR. SULSTON: You mean you're allowing for the uneven representation?

MR. CARRANO: That's correct. We've taken basically an 80 million base pair segment of DNA, synthesized it on the computer and gone through and cloned it; and the problems that come in, while you cannot get complete closure by the computer simulations, basically, you cannot clone every fragment. That's where it comes out. That's where the lack of closure comes in.

MR. GILBERT: Yeah, but see -- I'm struck by that. You're using a much finer mapping procedure than John is. So you're putting more effort in a certain way into the mapping procedure, and yet you're assuming finally a very low figure for the number of contigs you're gonig to get out of it; because 80 percent coverage -- you've got maybe two or three in your contig link and maybe two or three, at most four, cosmids. You have very, very short pieces, and you haven't faced the problem, well, what do you do about the rest.

So why are you planning to engage in that particular mapping approach?

MR. CARRANO: Oh, we have faced the problem what we do about the rest. The problem of doing the rest is basically the similar procedure that John has been using, and that is to use a series of fragments in yeast and using the end cosmid clones, the same approach Kahara used which is an approach similar to John's, use the end clones from each contig to be able to do rapid walking using yeast fragments or Not 1 fragments.

MR. GILBERT: You have a more uniform cosmid representation, you would have a much better -- For your two years of effort would have done much better for you. So the proposal as you're describing it is actually one where you spend a lot of effort constructing a few cosmid banks by sorting chromosomes, and you're very limited with DNA that's given. Therefore, you're not going to even contemplate constructing five or ten cosmid banks to give you a better representation going into it.

MR. CARRANO: I don't think that's true. We're at the stae of the art with cosmid technology that we can be right now, and I think John has stated that.

There are problems with cosmids, and it's not clear we can resolve those problems. There are problems of clonability. There's problems of representation, and we don't know how to get around that.

My estimate is what John said, that the lorus series of vectors probably are the best set right now to attack those problems; but there are still problems remaining. And that can probably get us up to 90 percent closure. We have to rely on additional methods to do the other. We may have to go to another set of cosmids to get overlapping libraries. We may have to go to yeast, but those will be done.

MR. McKENNEY: Does anybody have an estimate for what that last ten or twenty percent will cost?

MR. SULSTON: It's very difficult. I mean, I -- You know, as I say, if the yeast clones live up to their expectation, then another two person years, which is equal to what it would now take us to get back to ninety percent. But it jsut depends how they pan out. I mean, it seems they ought to, but --

CHAIRMAN BERG: This is now for a new total or--?

MR. SULSTON: I mean, what I'm saying, just multiply out so you can compare them, I'm saying we've taken two person years which is to get to where we are with our method.

CHAIRMAN BERG: So we're talking now maybe up around over 200 man years to get a fully overlapping cosmid map, if we add those two numbers together?

MR. SULSTON: Yes. I mean, I suspect that you can't speed up this last. It's just as Helen said, when you start fiddling around and having to go back and think about things --

MR. PEARSON: Two comments. One is that there are intermediate vectors being built based on bacteriophage P-1 that have the attraction of including 100 kb pieces of DNA; and these could be particularly useful as intermediate between yeast and current lambda base cosmids, and they could have a significant impact both on closure and on the clonability problem. But I'd like to reiterate a point that Keith mentioned in passing, and I know others here probably have better information than I.

That is, in doing the coli map in coli, there are gaps, and I know of at least two labs in America working on the same problem, one of which has the same gaps. So the issue of nonclonability for whatever reason is a potentially very serious problem, and we shouldn't lose sight of that in trying to say, yes, we can clone DNA.

CHAIRMAN BERG: Can I -- There's two other items on here which perhaps we can get some data on, and that is we're going to develop a lot of clones, both lambda and cosmid clones; and I assume that they're not going to just stay in your personal freezers. They are going to be distributed, and there will be repository, and perhaps we can get some estimate of what the magnitude of the problem is of (a) collecting, keeping and dispersing and keeping the information on such repositories would be. Can you help us?

MR. STEVENSON: I'll try. The interesting thing is this is the opposite way about the things usually happen. Usually, the ATCC gets materials after they've been thoroughly studied and researched, not at the front end of things. So you're going to generate a lot of material, some of which may be useful, and some of which may not.

At the present time, based on what we are currently doing in the contract with NIH, the cost to accession, work-up, quality control, prepare a small amount of material for distribution and to catalog a clone is about $1500 per. Now I should point out that what you're talking about, if the numbers here I'm translating correctly -- we currently have in the ATCC 42,000 items in inventory representing species.

It's taken us sixty years to get this amount of material. You're talking about somewhere in the neighborhood of a minimum of 350,000 items in maybe three or four years time. This is way beyond the state of our art at the present time to contemplate doing that in an orderly fashion and making all this stuff instantly available to the scientific community.

There's nothing that is mysterious about the process, but certainly, to handle this kind of material and do it efficiently, we would have to do such things as improvement of our accessioning and materials handling stuff, using bar coding, a variety of things.

When we get down to the production of this material and freeze drying it so that it would be availabl for distribution in the mail at minimal cost and not as frozen material, it has to all be hand done. We can't put this stuff into bulk lyophilizing apparatus, because there's cross-contamination, and so you would wind up with mixed clones. It has to be done in a manifold or individual type preparation, which is probably going to be the limiting factor in this unless we can figure out a way to automate that.

Currently, we have no incentive or no funding to do that, simply because the problem has never come up before.

CHAIRMAN BERG: Is that the proper number, 350,000 clones? I'm just trying to figure out, from the number of cosmids plus the number of subclones, let's say, and lambda clones -- Is that on the low side?

MR. McKENNEY: Well, if one takes the Kahara data for what it is -- he generated 4,000 clones for coli, and that reduced to 481 clones to cover the whole genome; and simply multiply that times -- instead of 4.5 or five times the bases for coli --

CHAIRMAN BERG: So it's a thousand times more.

MR. CARRANO: Well, if I've done my calculations right here, it's 10^5 cosmid clones with an average insert of 40 kb will cover the human genome one time over. So one is going to need at least five to tenfold. So you're talking of --

MR. SULSTON: I think a twofold depth is reasonable, so that you have --

CHAIRMAN BERG: But that's just for the cosmids. That may be -- The question is, is one going to want to have subclonings of cosmids or of lambda clones, or small segments in the cosmids?

MR. GILBERT: I think not of the distribution. A goal of this sort of mapping is to create a reference bank of clones that would be available for distribution. That reference bank -- Tony's figure -- That's 100,000 clones. As you would go through and construct those maps, my estimate has been that one would run through something like 500,000 clones.

CHAIRMAN BERG: To construct it.

MR. GILBERT: To construct, right. And you might lay down -- I mean, one strategy is, of course, to lay down those 500,000 clones as you pick them in the first place, then go back and distribute from that.

An alternative strategy is don't lay down those, but take each of the 100,000 clones separately, grow it up and multiply it. I think that's what Dr. Stevenson was thinking about. I think that's more expensive than laying down the clones originally.

MR. SULSTON: Yes, I agree. I think it's less attractive. I think what you want to do is to keep everything. Let me just say what we're doing, because I think our problems in miniature are just the same.

We're growing in a way -- using the flow "micronic" system, so that things end up in little 1 mil. test tubes, and most of those things are never going to come out of the Minus 70 freezer again. But when it's of interest to somebody, one may go and pull out indeed six or ten of these, if it's available, to send to a particular lab that wants to make restriction map or do some sequencing, whatever.

So presumably it's more practical in planning this to have a start that's available to do the freeze drying and so on as a particular region is called for, which, it seems to me, even in the human that you're not going to have an even request. You've going to have some popular areas. And at that point you would do your lyophilizing and so on so that you can send that out multiple times.

MR. STEVENSON: It's not efficient.

MR. SULSTON: Is it not efficient to do that?

MR. STEVENSON: No. I don't disagree with the fact that you would want to save a lot of this material as seed stock in liquid nitrogen -- temperature. I think that is fine. There is -- And we can develop costs for that, but the cost of putting something down like that and keeping the records on it is not trivial. I mean, it's something you have to do. But then to bring it out and to make something from that for one person means that there is an awful lot of hand labor, an awful lot of gearing up just to do that; and it's not a very efficient way to do it.

Also, when you have that material in, you don't like to go back and repeatedly freeze and thaw, freeze and thaw, to get out your seed material for making a batch.

MR. SULSTON: Oh, no, you musn't do that. You take it from the frozen --

MR. STEVENSON: Well, all right. The point is that, while you're putting it down in the first place, you may as well make a small amount, say, 10 amps or something like that, to put in so that you don't have to go through that process.

MR. SULSTON: Well, yes, I entirely see your point. And of course, what you're saying is right for your 4200 items. But what I'm saying is -- I mean, I wonder if you may not find that int he end the cost will be lower and the service better if you keep everything in seed initially and just respond to the market, so that if people want it they can get a lot of clones from one area, and then the neglected areas can be maybe never thawed.

MR. STEVENSON: Well, it depends on the approach. The scientific community that owns our organization has taken the attitude that they want to pay as little as possible for the culture that's distributed to them. Our accessioning costs for a lot of this material represent what the businessman would call capital investment. In many cases, the government has paid for the capital investment in the stock, and subsidizes then the distribution by putting money into the place so that we can keep the prices of cultures down.

The government may balk at that in a project of this size. So that if one did in fact have the seed stock and charged what was really the cost for the individual ordering a sample, yes, it would be certainly doable; but the point is, given the present policies and the attitude of the scientific community, it would be terribly hard to try to operate. We would go broke rather quickly in trying to provide cultures at the standard price using the type of system that you're suggesting.

MR. SULSTON: Well, I suppose -- In fact, it may be a compromise solution, but it's the right way to go, that you would have available all the seed stocks which could be taken out at a price, but that there would be at some point when the map was more or less complete a canonical set prepared. In fact, this is exactly what we'll end up doing for the nematode, a canonical set prepared which will be, as you say, put away multiply; and those are the ones that you'll send out.

CHAIRMAN BERG: John, you're saying, there are two expense items. One is to keep all the seed clones, all the original ones, stored; and maybe 100,000 pounds of which are reproduced for distribution.

MR. SULSTON: Yes. I think you've not only forgotten the great value of the high depth seed clones. I made the comment that twofold would be enough.

CHAIRMAN BERG: So what I heard -- I'm trying to get to this number. We're talking about storing and distribution, -- a capability for storing and distributing 100,000 clones, and storing 500,000 of the original clones. Is that the number? That's what I thought I heard?

MR. COOK-DEEGAN: Now is there any way to separate out those costs, figure out the minimum cost for just dumping something i --

MR. STEVENSON: Only if one determines a policy on how they would be paid for, and whether you were generating these things on a custom order one at a time or whether you were making them by the batch.

CHAIRMAN BERG: Well, I thought we decided that, roughly, having the ability to distribute 100,000 clones on demand, without having to go back into the thing, and maintaining the seed clones as a back-up -- Is that -- That's what I thought we were talking about. And -- whoever pays for it. What would that be? Why don't you think about it. We'll come back to that.

In data bases, which is clearly a lot of stored information on this, I haven't heard anybody come up with a -- I mean, obviously, individually, Los Alamos has got its own way of keeping track of its cosmid clones, and John, I assume, at the MRC you people have a good inventory; but as the number grows, when you start talking about 5,000 clones, what would it take, and could that be done within the same facility that's going to deposit them?

MR. SULSTON: I -- The computing?

CHAIRMAN BERG: Yes.

MR. SULSTON: Yes. I didn't mention that, but it turns out to be trivial; and I think it would still be trivial when applied to the human genome compared to the other costs. The only thing that has to be done is to develop an appropriate parallel processing device, and that's cheap.

CHAIRMAN BERG: So what you're saying is that all of this matching and lining up of cosmids doesn't really require a new generation of computing capability.

MR. SULSTON: Do you object to that? That's my conclusion.

CHAIRMAN BERG: Is that fair enough?

MR. HOOD: Yeah. I mean, I think you can divise engines that can do it much faster, but it's --

MR. SULSTON: Yes, and it would be worthwhile in the context of this.

MR. HOOD: We know what the tools are.

CHAIRMAN BERG: So is the number that's going to go in there small relative to the rest of the total?

MR. SULSTON: Yes.

MR. BURKS: I mean, I don't know -- I agree completely with the idea that the tools or the technologies are available, but I don't know that there's a trivial cost associated with thinking of the best way to implement the tools for the various data sets relative to mapping and sequencing and the other aspects here, and how to connect them, and how to make sure that people in all different locations have access to them.

CHAIRMAN BERG: Is the magnitude of that number likely to fall just in the ranges we're going to be able to establish anyway? I mean, if we're talking about $25-$30 million -- Isn't that what we're talking about?

MR. HOOD: The way you have to look at the computing is you've got four problems. One is the acquisition of data into a data base. The second -- and this is both mapping and sequencing now. The second is, thinking about how you're going to structure the data base, you may have 15 billion units that you have to sort through. Then the third is pattern search, and some of the pattern searches will be much more complicated. Then you get into sequence questions. Then the fourth is dispersal.

I would agree with Christian. If you look at it in that way, it -- although we may know the general tools we can apply, it's going to cost a lot of money to figure out how to do it.

MR. SULSTON: Well, yes, I didn't mean to trivialize the task. I mean, I think that the project deserves the very, very best computer people coming into it. But I would still have thought that the cost of doing that are low, compared with the whole thing.

MR. BURKS: I -- We're getting late for the break, but at some point I've done some numbers on computing costs.

CHAIRMAN BERG: Well, we're going to come back, because each of these categories down here, clearly, under sequencing have a very -- I was trying to sort of just focus on the basis of what we've heard that for obtaining the libraries, making the cosmids, keeping track of them, the overlaps, determining their order, and the kind of storage in a facility where they're going to be made available for, that -- what's the number? Okay. You're trying to compute --

MR. HOOD: I think we've both thought about it in terms of the global aspect.

CHAIRMAN BERG: So is this going to be buried in that cost?

MR. HOOD: I think it's going to be a very small part of that cost.

MR. GILBERT: I think the data base is buried in it. The handling, the bar coding, the keeping track of 100,000 objects and knowing where they are and that sort of thing. But that's part of the --

CHAIRMAN BERG: Yes. That's their job. Okay. So under this -- Under physical map data bases --

MR. STEVENSON: The cost to store 500,000 clones for thirty years at nitrogen temperatures would be about $250 million, and the cost to produce 100,000 clones for distribution would be around $10 million.

CHAIRMAN BERG: You mean $10 million per year?

MR. STEVENSON: Well, however -- you know. You don't produce them unless you need them. You want to figure on what it would cost to produce 100,000 clones. So, you know, whatever rate -- The problem is, if this is something that generates an awful lot of interest very quick and a lot of people do it, and this is not an orderly accession, then I'm going to have a logistical problem. But basically, the capital investment is in getting refrigerators. We would need somewhere in the neighborhood of twelve of the largest liquid nitrogen refrigerators currently available to store one each of these 500,000. And depending upcn what your number would be for a reliable number of replicas to have for the seed, you just take that number times those refrigerators; and we're talking about a large tank farm.

MR. SULSTON: That's an interesting point. you see, I think with something like this, it pays you not to have replicas but to treat the thing stochastically and say, if we lose one, too bad. We just try and minimize the possibility of that. With 100,000 across there, you have the replicas. For the rest, you do your best. And our own approach is we don't have replicas now. We have a single frozen tube, and you take cells from that without thawing it.

I know it's not the way you think and, of course, (laughter) I would probably be horrified if I sent you a special thing and you treated it that way. (laughter) If you've got 500,000 things, then I think it pays to tell people and do it in a different way.

MR. STEVENSON: I think I'd want to move to Ft. Knox. (laughter) Well, currently when we get seed stock in, we store it in two additional places to the home base, because it's too precious. There's too much invested in it to run the risk of losing it.

MR. SULSTON: I suppose two places are sensible, but I think you don't have to --

MR. STEVENSON: But when you're doing this, you know, the costs keep going up. Currently, I have twenty-four tanks in the repository. You're talking about taking half again just for one each of these. So I've got to think about a new building already, you know; and these things start pyramiding, and it gets me very nervous.

CHAIRMAN BERG: Let me see if I can sum up here so we can take a break. In the production of clones, you estimated of the order of $8-$10 million. In the ordering them, you're talking about maybe $10-$20 million. Then in repository we're talking about a quarter of a billion dollars, and data bases just sort of trivial in that context, if I understand correctly. So we're talking about roughly $200-$300 million for that category to obtain a physical map and store all the clones and be capable of distributing them on demand.

MR. GILBERT: You have to look at the storage cost again, because at that point it's worth making the whole thing over again every two years rather than storing it. You think, you're estimating $20 million to do the map the first time. The second time is going to be a lot easier, because -- So the actual cost of making that map -- the value of that map --

CHAIRMAN BERG: So what you're saying is when somebody wants a clone, you send them the protocol instead of the clone. Protocols you don't need cold storage for. I'm astonished by the number.

MR. SULSTON: Can I ask one more question. You mentioned -- Again, bear in mind that these are not so precious. What's the cost to store them in electric core refrigerators? Is that cheaper or not? At minus 70, I mean.

MR. STEVENSON: The cost of storing them at minus 70 is no that much different than storing them in liquid nitrogen.

MR. SULSTON: Okay. That was my question.

CHAIRMAN BERG: Jim?

MR. WATSON: The ten to twenty just seems --

CHAIRMAN BERG: It depends on the cloning. It depends on filling in gaps. So I was doubling it. I mean, the original estimate --

MR. CARRANO: I think Jim is correct. I think we came up with 200 person years and, if we apply a factor of 150 k, that's $30 million.

CHAIRMAN BERG: Okay. Sorry.

MR. WATSON: So it might be $50 million. I think you would really be misleading people to say less.

CHAIRMAN BERG: Okay. Sorry. So you mean all the way up to getting an order library, $50 million, and 250 we're talking about of possible storage costs.

MR. STEVENSON: That's over thirty years.

MR. WATSON: Most of that would be, I presume, to buy the refrigerators.

MR. STEVENSON: There's not much point in doing it if you're not going to do it for the long haul.

MR. WATSON: No, but what's the initial cost going to cost?

MR. STEVENSON: That's where the heaviest part is, the preparation and the data management. The cost per year would be -- per ampule would be in the order of fifty cents to a dollar per ampule, just for the nitrogen and the facilities.

MR. WATSON: So your investment over the first five years would be what? Say, ten years? Almost the same as it is for thirty years?

MR. STEVENSON: So we're talking about $50 million a year for five years or something like that, something on that order, except you have to build the building, too.

CHAIRMAN BERG: It's a big number, and it's up front. I mean, most of it is up front. Okay. I think we could use a break.

(WHEREUPON, the workshop recessed for lunch at 12:00 Noon.)

- - -

AFTERNOON SESSION

Time: 1:22 p.m.

CHAIRMAN BERG: I want to take a few minutes to go over some ground that we had already covered, and I think I'm going to ask John Sulston to speak to the issue of the physical map; because in the earlier discussion we were -- The estimates we made and the strategy which w were discussing was essentially chromosome by chromosome cosmid maps, overlapping those cosmids, and creating a physical arrangement.

I think John has pointed out and we've already discussed that the advent of cloning in yeast segments of the order of 400 kilobase in length rather than being restricted to forty really does change quantitatively what kind of estimate you can make.

It also impinges on the cost estimates for storage because, clearly, if you have to store only one-tenth as many clones, that has a very significant impact as well. So I want John to just address that, because I think in the conversations I had with him when we broke up, it seems that the numbers we were talking about might be very drastically altered by assuming that one is going to go now directly to making a yeast YAC liabrary and using that as the basis for the physical map rather than in cosmid libraries at one-tenth the size.

John, you want to --?

MR. SULSTON: Well, I think you said it very well, Paul. (laughter)

CHAIRMAN BERG: Well, come on. I'm just translating from British into American. You go ahead, please.

MR. SULSTON: Well, I mean, I think that's an admirable summary. I mean, if Maynard Olson were here, then I think he would be talking about doing exactly that.

MR. COOK-DEEGAN: He's in Europe.

MR. SULSTON: I know. I know he's thinking along these lines. You see, when you think about it, if one clones into YACs of a mean size of, let us say, 400 kb, which is current technology, not projected, and then one takes the trouble to devise a rapid, efficient fingerprinting or some -- to generalize here, I'll just say a matching procedure for those YAC clones, then by handling, let us say, four times as many clones as we've done with the nematode so far, four times as many YAC clones, you would end up with the sort of coverage that we have, or better perhaps, if the YACs are more random than the cosmids are. One might, in fact, have a very complete map at that point.

Now, of course, it's a low resolution map in the sense that each clone is going to be 400 kilobases long. On the other hand, the overlaps in the way that we were discussing, putting them out in depth, the overlapping nature of these clones will allow one to precisely map, for example, the position of a particular RFLP by hybridization. And so, although -- I mean, I gave you -- Wanting to stick to the present tense, I gave you projection of the state of the nematode, I seriously think that I would follow Maynard if I were to start now in the human and simply go for the total YAC first. Then at the same time, to consider how that should be related to cosmids and/or lambdas.

CHAIRMAN BERG: The way you described it to me, I think the direct point which struck me was you bypass the whole business about chromosome sorting. You wouldn't go for chromosome specific libraries.

MR. SULSTON: If you were going --

CHAIRMAN BERG: You would be making a general library right away.

MR. SULSTON: I think, if you're going into YACs, then it's probably fair to say that you would not start with sorted chromosomes because of the difficulty of getting very long pieces out. And so it would provide, in the context of a single lab, bear in mind -- this is not an institutional project now; this is a small group project, to do the total YAC map of the human.

This would be -- This would mean that you would approach it directly. That's right -- without sorting the chromosomes.

Now how you fit in the other things, I don't know. I mean, that's what we should now discuss.

MR. CAHILL: What's a YAC size, one, three, four million, for those of us not quite in the field.

MR. SULSTON: Sorry. No, it's -- I was saying, it's an order of magnitude more. The cosmid is forty. The YAC is -- It's reasonable now in present technology to make a YAC library of mean size 400 kilobase.

MR. CAHILL: Four hundred? I thought it was bigger than that.

MR. SULSTON: Well, Maynard quoted as having made one of 900. You see, the problem is, I think, getting that very big DNA like 8 to 10N. But why don't we stick to what we know we can do, which is to make a YAC library of mean size 400. That then, you see, brings the -- It gives you an order of magnitude advantage, arithmetically; and that's why I say, a further factor of four of the nematode would bring you up to the total human genome.

That's why I say it's a project of a small group, to take that number, the appropriate number of YAC clones, fingerprint them or whatever you want to do, and match them together. AT that point you might have a complete map at that level. And the YACs are very suitable for storage, as far as we know. I mean, certainly, they can be stored in exactly the same way as cosmids. They're very --just stick them in the freezer.

I don't know if you can lyophilize them or not. Not sure about that. The disadvantage possibly is that, in order to get your fragments out, you have to run a pulse field gel; but that's not such a big deal. We routinely are running pulse field gels with twenty lanes, purifying twenty of these fragments. One could improve on that.

We can go into science fiction about possible future ways of doing it, but at the moment that's the technology for taking out these individual chromosomes.

CHAIRMAN BERG: Wally, you wanted to add something? Say something?

MR. GILBERT: Yeah. I think some of the strategies for a library is very much connected into what you would plan to do with the library afterwards. But my own view, since I think one of the major purposes of having a complete physical library is to be able to go on to sequence the DNA, I would think that the library that one must create is a library in which you can go immediately to sequence the pieces of it.

Therefore, one would actually try to ideally construct either a lambda library or a cosmid library, something whose peak sizes would lend itself to the sequencing. I would think -- I do thinkit's an advantage to construct that library from the basic material, the reference DNA directly, and then to order it by other techniques, which just got mentioned earlier.

I do agree probably with a statement I think you were suggesting, that I think it's advantageous not to separate the chromosomes, that if one were attacking the problem in a single site that one would take bulk DNA, make many, many libraries, cosmid libraries, out of it, map those with respect to each other, and then use something like the yeast clones as part of the ordering -- of the ordering strategy. But the yeast library, while an ordering library, I don't think is going to be a reference library in the form that you would want to take those clones and subclone them.

MR. SULSTON: Why not?

MR. GILBERT: Because that gets you tied up again totally in the stability question.

MR. SULSTON: Well, yes. It's not clear to me at this point --

MR. GILBERT: You have no way of sort of referencing the stability in the way of testing it.

MR. SULSTON: But it's not clear to me at this point that the YACs aren't just as good as cosmids, maybe better. We don't know that. It may prove -- They may prove to be better than cosmids.

MR. COOK-DEEGAN: Wouldn't it be a prudent policy to do both? Would it be useful to have both cosmids --

MR. GILBERT: Yes. As I was sketching it, one would do both. The cost that we're estimating here for the cosmid libraries are costs of the order of tens of millions, not assuming that there are improvements in technique as one goes through. The costs that John is suggesting for the YAC library is, of course, single millions or less. So that it's not -- If your question is what is the goal, I think the goal is a library that maybe we can go on to sequencing; and that library can be achieved at costs in the order of a few tens of millions of dollars.

I think the discussion we've had on storage has to be rethought, in the sense that it does -- you would not choose to store that library at a cost much above the construction cost. And, therefore, you would seek methods of storing it that are comparable to the construction cost, and those methods are either storing the DNA, storing lyophilized cells, setting up because you have a problem of storing five million lyophilized vials, setting up a way of constructing those lyophilized vials at a cost again comparable to the cost of the library.

MR. SULSTON: Yes. So I think one important thing about the YAC library is that, for many of the users, I think, the type culture collection the YAC clones will be perfectly suitable. This is for genetic medical uses, I think.

It might fit in very, very well with the RFLPs. The contigs made from YACs would bridge nicely between the RFLPs.

MR. HOOD: I think it would be preferable.

MR. SULSTON: And it might be that the sequencing application and the detail molecular application would not need to go to the type culture collection. This will be restricted to a relatively small number of laboratories who would exchange their clones on personally arranged collaboration.

MR. HOOD: I think one of the things that will happen is that we'll automate cosmid or whatever, and then if you store it as a 1 megabase piece, you could convert it into cosmids or lambdas or whatever you want for looking at specific genes. That does cut down the storage problem enormously, and I think that's a really important point.

CHAIRMAN BERG: Part of the reason for bringing it up was to really put this cost estimate into perspective. As Wally would say, it doesn't make sense to be spending an enormous amount of money which is ten times the amount it costs you to make it.

MR. HOOD: So I'd say one way we can cut down the cost is to store much bigger fragments. I think that's an important point.

CHAIRMAN BERG: Yes.

MR. HOOD: I think the second point is, I think when we think about it, the technology is going to really facilitate automatic storing of large numbers of these things once again. So I think those two things together will bring it down to within a reasonable range.

CHAIRMAN BERG: So the number we came up with at the end of the morning session essentially is like almost an upper limit, and maybe by a factor of two or three actually lower on the basis of the considerations that have been made here. That's what I'm drawing from this --

MR. HOOD: The $60 million figure.

CHAIRMAN BERG: Yes.

MR. WATSON: John, if you went from the YACs to the cosmids, you're going to still have to order the cosmids within the YAC.

MR. SULSTON: Yes, that's right. Well, that's why I -- One just has to go off and do some experiments and try out some things.

MR. WATSON: And figure out how much more it would cost.

MR. SULSTON: See, the thing is, I wanted to say, is that there are great advantages to having this relatively low resolution physical map, just like the RFLP one, and then the other -- The sequencing labs can take care of the problem of going to the low resolution map to the sequencing whatever way they see fit. So presumably, they'll want to do experiments and see how to go.

One can either probe --

MR. GILBERT: Wait a minute, John. You're promising $2 million on the YAC maps and letting the sequencing labs take all the rest of the problem. I'm proposing something very different, say, $60 million and getting both maps.

MR. SULSTON: I'm just saying, I don't know the most efficient way of going from one to the other. I mean, do you really?

CHAIRMAN BERG: Okay. No, let's leave it at that.

MR. GILBERT: I think the strategy, the one that you're following now, is very efficient.

MR. SULSTON: It may be that it's best.

MR. GILBERT: Because you already have contigs.

MR. SULSTON: Yes.

MR. GILBERT: And rather than trying to exhaust yourself closing it, you will now take the YAC map and --

MR. SULSTON: That's right. That's right.

MR. GILBERT: You know, had you made the YAC map, you would still have the same problem. You have to make a set of contigs and order them. I don't think you would choose to order your clones --

MR. SULSTON: The thing is -- I think the thing is that it would be reasonable, that the YAC map has justification of itself for medical genetic uses for firmly tying together all the RFLPs and offering you a probe for any point on the human genome.

MR. GILBERT: Depending on the resolution.

MR. SULSTON: Sorry? Depending on the resolution you want, that's right. But bear in mind that it will be relatively straightforward to remove particular fragments. I mean, you can, for example, have a not restriction map or something, so you composed on that and define fragments that way. You know, you can create, if you like --

MR. GILBERT: It gives you everything except easy access to large amounts of DNA.

CHAIRMAN BERG: The oother thing it does, it seems to me, it gives you another approach, an independent organism for avoiding the problem of not being able to clone certain regions or getting overlaps. If you go for 400 kb fragments in yeast, you may in fact be able to clone the pieces that you can't clone in coli, and certainly going in bigger pieces allows you to get --

MR. GILBERT: Be careful. We already have a map of separated human chromosomes in CHO cells, and you have a map of fragments of human chromosomes carried in animal cells. And both of those have advantages, and they're easy to store and there are small numbers of them. And they all have disadvantages we could also go through listing, including questions of stability of chromosomes, is the thing really there, and all of that.

You're going to have those disadvantages and advantages at every level. So I don't think there's an automatic, oh, there's a wonderful solution to the problem. It just happens to be at this level.

CHAIRMAN BERG: Okay. Let's leave the number around sixty then. Okay? So, I mean, essentially, all approaches would be included within that kind of number. Okay.

The only thing that might change was in fact the storage costs, becasue if in fact you agreed to store them as the large segments, then the number of things that have to be stored is markedly reduced.

Right? So that was one number which would --

MR. COOK-DEEGAN: Just for our purposes, to follow that point, the way Wally would solve that problem if he were writing the section of the chapter, sounds like he would say -- the rule of thumb you use is you don't want to spend more to store than it costs to construct in the first place. Is there consensus that that's a logical rule of thumb for deciding what the criterion is for how much is going to get stored? Is that --?

CHAIRMAN BERG: I think that's almost been the actual practice of most people who collected large numbers of mutants and phage mutants. In the end, once it's described how they were isolated, it's probably easier to go back and get them than it is to store all of them.

MR. COOK-DEEGAN: That's a nice rule of thumb. So that allows us to write that section.

CHAIRMAN BERG: Okay. Let's go on to technological developments. Sequencing. I'm not sure exactly how we're supposed to come up with --

MR. COOK-DEEGAN: This is where defining the vision is crucial.

CHAIRMAN BERG: So, Lee, why don't we let you start.

MR. HOOD: Okay. Well, let me make a couple of brief points first. I think in this project it's important to keep in mind that one major aspect of it will be technology development as opposed to production line operations. I think the technology development is critical to every one of the aspects of it.

The way I look at it, it falls into four categories. One is DNA handling. We would like to make things easier and more efficient for generating libraries and isolating clones and amplifying them and preparing them for mapping and preparing them for sequencing and so forth.

The second is automating the various kinds of mapping techniques, and there are a variety of different ways that we can think about doing that.

The third is automating sequencing itself, and that means all the way through from creation of these sets of fragments. We want to automate pouring gels. You want to automate loading samples on gels, and we want to automate the first instruments.

Then, finally, there is the whole aspect of data analysis, which is this acquisition and storage of analysis and then distribution.

My own view is this technology development should be a very significant fraction of the total cost of the whole human genome project, because it will amortize very quickly in the speed and the accuracy and the way in which things are going to be done.

I think the second thing that's important to remember is, if we're thinking about an endeavor like this, we aren't just going to take pre-existing groups and make sure they get funded as well as they have been or perhaps a little bit better, but we have to really think in terms of scaling up and adding maybe in the first five year period ten or fifteen groups and then successively additional groups. I would guess that we probably want to add them at kind of a modest intermediate size, maybe one ro two people up to perhaps ten people.

These groups could do -- could be associated with any one or more of these various aspects, be it mapping, be it sequencing, be it any one of these aspects of technology development. In fact, presumably they would combine together these things, so you would have people doing real mapping and sequencing in conjunction with the technology development.

This brings us, of course, to the point of, if you're going to bring in new people and set up new groups, the capital costs are going to be really significant. It will be a very significant fraction of the cost of this project.

You've got to build places for these people to carry out the science, and you've got to set up the kind of instrumentation that you need.

So having made those general comments, I'll talk about sequencing, where we stand, if you wish.

Today it looks like there are basically three major approaches to DNA sequencing, and they're defined with regard to the analytic end. All of the methodologies use the production of DNA fragments that are successively a nucleotide longer, and you can do that by chemical methods or by enzymatic methods. But what you have to be able to do in the end is to analyze DNA fragments that differ by a single nucleotide from one another, and to somehow identify which nucleotide they differ.

So the three approaches are, on the one hand, using a single label and taking a picture in the media that you use to separate these DNA fragments. That's the classic -- Radiotography is the class way of doing this.

Of course, one of the recent very interesting developments in this regard has been the multiplex sequencing approach of George Church. Here the idea is, you can put ten or twenty or fiftn sequencing reactions down on a set of four lanes, and then use different probes successively to analyze the sequences that are present in those four lanes. What this does is give you an enormous economy of running gels and blotting and whatnot, and George feels, I know, that this kind of multiplex sequencing is going to permit one to think about doing megabase a week sequencing or so forth.

The critical point with regard to this -- these procedures, I think, on the one hand are: Can ordinary people do them? George is technically extremely good. And on the other hand, can one automate the gel reading to give an appropriate accuracy? I think we don't know about either of those things right now.

I'll just say, my own bias is that four lane analysis is going to be rather difficult, because -- at least the way sequencing is practiced now, because of nonhomogeneities in the gel, differences in the lanes, and the technical difficulties of getting the computer to match up them so you can call up the sequencing. And that's a limtiation in the second method, which is again a single label method, often a flourescent label; but the analysis is done in real time. So you'll carry out your four sets of sequencing reactions, terminating at the four bases.

You'll put them each in four lanes. You'll code them with a flourescent -- the same flourescent molecule flourescine or whatnot; and then you'll have a detector somewhere down in the gel to identify them as they migrate by in each of those four lanes.

There are several groups that have developed this kind of technology, and there's no information available that I know of on automated analysis. But I think there will be a nontrivial problem in being able to lead four lanes and standardize them and get the very high levels of accuracy we need.

Then, of course, the third approach is using four different labels, real time analysis, and the analysis will be done in a single lane. This is the instrument that AVI has, and it's an instrument of the type duPont has recently developed. Here the advantage is each lane uses -- can use the four labels to serve as an internal standard.

Now I think probably there's no area in which the costs are going to change more dramatically than in an automation of DNA sequencing, but I can give you kind of what people guess are the state of the art costs right now.

In typical companies I think estimates are that DNA sequencing costs somewhere between $1.00 and #1.00 a base. So if you wanted to do the haploid human genome, you would be talking about $3 billion to a first approximation. Now the Japanese have put together a consortium for automating really the major aspects of DNA sequencing; that is, the construction of the fragments, the pouring of the gels and the analysis of the gels. They have actually used several of these different analytic gel detection techniques, but their estimates, I guess, now are of the order of seventeen cents per base. So that brings down by a factor of five what the cost might be.

The current AVI machine, if you start with your DNA clones in M-13 single stranded vector ready for sequencing, the estimated cost there under optimal conditions might be of the order of six to eight cents per base. And it's clear, within the next six months that the technology for bringing that down to maybe a penny a base or two pennies a base is at hand.

So that's the range of things that one can do right now, and it talks only about the costs of sequencing, given the DNA, and given the DNA in states that appropriating for sequencing. We're not talking about starting with a YAC and going to a cosmid and going to M-13.

CHAIRMAN BERG: One point I just want to clarify. Some of these estimates are given in terms of base pairs. Some of them are given per base. Clearly, if it's per base, it's a factor of two.

MR. HOOD: I think in all the cases that I talked about, I meant to say per base. It's talking about per single strand reaction.

CHAIRMAN BERG: The second question is: How many times do you have to sequence the DNA to be reasonable confident that you have a correct sequence? So in other words, it's not three billion base pairs. Maybe it's the equivalent of 15 billion base pairs, or ten or more. You're saying more. Let's say, fifteen to twenty billion base pairs. So in fact the cost, the unit cost is going down, as I understand it from what I first heard of $3 billion, base pair of $3 billion.

Then now you're tlaking about 20 billion base pairs at whatever the unit cost turns out to be. Am I right, or am I --?

MR. HOOD: Well, I would guess that the factor probably is going to be three when we get finished. And I would say, because you'll have comparisons on both strands to analyze things against, and you'll be able to build into the software program the intelligence of the sophisticated sequencer where he can say this strand is right because on the other strand we've run into compression.

So you will be able to cull out in all those cases of ambiguity that the probability is 98 percent that it's correct here. Then if you had a third run at that, you could probably unambiguously separate those things out.

So I think with good software or culling, with error rates that are, hopefully, eventually somewhere between .5 percent and .1 percent, I think you would be able to do it with --

CHAIRMAN BERG: With ten?

MR. HOOD: -- with -- No. I think with three. If you get a good physical map -- Well, I guess we're getting now to a question of what's your strategy for sequencing. Okay.

On the one hand -- Well, there are two basic strategies. One is the shotgun method where you do random cloning and, just as for mapping, if you sequence each of these random clones you'll have to sequence some ten or fifteen times to get maybe a 90 or 95 percent closure.

Now I can envision strategies where you won't have to use the random approach at all, but you use a directive approach from the very beginning. For example, let's suppose that we have very, very sophisticated automated techniques for mapping. We could thinking about taking our set of cosmid clones and constructing detailed six-hitter maps, and then detailed four-hitter maps; and then you could generate a random set of fragments which could very quickly be mapped onto those detailed maps. Then you could pick out exactly the set of fragments you wanted to sequence, so that you could cut down what you had to do by sequencing.

The critical question there is whether sequencing costs significantly more than mapping those or not. Those are technical issues that we can discuss. Or you could think about very powerful strategies for actually being able to start with a cosmid and do -- label it at one end, and do partials where you could create that successive 500 base pair deletions. The whole nested set of fragments would go all the way back, and then have analytic techniques for separating those fragments, which could then be sequenced at the end to give you --

So those are the kinds of strategies you can think about, and I think in the future my own belief is we'll orient ourselves toward these directed methods that can be automated rather than the random methods.

CHAIRMAN BERG: Wally?

MR. GILBERT: I was going to -- I'm also the source of some of these numbers. I was going to say, my convention is to try to give figures for the ultimate double stranded product, because that's the thing we're interested in comparing. And the effort that it takes to get that double stranded, as he just pointed out, does depend on the strategy.

One of the problems, as people look at DNA sequencing today, is that a great deal of the effort is in the subcloning, identifying fragments in sequence, then going on sequencing. So lots of the estimates for how far and how much time it takes turn on DNA handling considerations rather than on the actual physical process of sequencing.

Thus the underlying strategy is very important. The existence of a physical map, let's say, a cosmid collection, changes the underlying strategy; because rather than sequencing DNA differently every time we come to the problem, the problem is going to be sequencing a 100,000 pieces of DNA that are the repetitive. It was the problem of the same one each time over.

That, I think, will lend itself almost immediately to a nested deletion approach, because again that's something which you can imagine essentially turning into a rote autoamtic process, and that will then provide you with a series of DNA fragments, each 500 bases different in length from the one before which you would then sequence in a again repetitive fashion.

In that case, you are going through about two to three times the final sequence in terms of the raw sequencing. The estimates -- I suppose I don't believe the Japanese estimate, because I saw the chart that Wata put up which led him to those numbers and that, I regarded, as pretty imaginary.

I do think that the current technology, the application of machines with multiplexing technology can produce grades in DNA sequencing that are of the order of a megabase per working person.

CHAIRMAN BERG: So this number of two cents a base -- does that suit you better thant he 17 cents?

MR. GILBERT: No, the underlying cost figure I used is about ten cents a base, final base. I think that's achievable with the current technology, and I would regard the critical thing to be demonstrated as one goes into such a project is that that is a reasonable cost to be achieved.

MR. WATSON: Does that include the nested deletions?

MR. GILBERT: Yes.

MR. HOOD: See, the estimates are very close, because finished sequence means you've done it three times. It's two cents a base. That's six cents plus the nested, the mapping; and that's certainly --

CHAIRMAN BERG: Oh, I see. Your cost is to get the sequence three times over, to get the whole duplex structure; and your number was --

MR. HOOD: No, no. My two cents a base in the next six months is single stranded, single sequence --

CHAIRMAN BERG: No, I understand. So you're very close to ending up with the same --. Anybody else?

MR. WATSON: Really, Wally, on the nested, is that optimism or that you really know you can do it?

CHAIRMAN BERG: Jim, would you speak up.

MR. GILBERT: He just asked me whether it's just optimism, but I would say it's optimism.

MR. WATSON: Okay. Because you could be wrong by a factor of four. That is, really, no one has shown that you can get them that fast yet.

MR. HOOD: Get them that fast? You mean --?

MR. WATSON: I mean the nested, that the cost would actually be that much. The running of the gels, you have a much better feeling as to where you can probably reduce the cost.

MR. GILBERT: Okay. Except I think that's a way of defining the problem. There are many ways of defining this particular problem. One is that we must have massive improvements in every way in which we do things. That lends to definite development in all possible ways.

Another is to say: Well, in fact, we think the current technology can do a certain thing; and you can say, well, let's demonstrate that. You know, let's demonstrate that the nested deletions in this approach will give us a cost of the order of ten cents a base on a suitable scale for doing it; and that then predicts a certain cost for the entire project, a way of providing a goal.

You can say do the project at a total cost of the order of $300 million, if we first demonstrate that we can do it at ten cents a base. The critical element there is to demonstrate that we can do the nested deletions in an appropriate automated fashion.

MR. COOK-DEEGAN: If I could build on top of what seems to be something of a -- on that figure. Thinking through planning on a project like this, Wally, you've gone on record before as saying the first step is to do first one percent, the second step is the next nine percent, the following step is the next ninety percent.

MR. GILBERT: Other way around.

MR. COOK-DEEGAN: So that's saying roughly equivalent cost for each of those -- again, these are rule of thumb.

MR. GILBERT: Yeah. When I first thought about this project several years ago, I actually said that in terms of thinking about the project over about a thirty year period and thinking of starting with technology which in fact I didn't realize was as good as it is today, and saying, well, at ten years one might do one percent and improve the techniques; and thinking of it broken up that way.

At the moment I think the techniques are much, much faster than I thought in 1984. Now we have the machines. We have a multiplexing technique, and I expect that the base technique we have is the one that can do ostensibly -- they can do a figure of the order of this megabase per person year. And at that level I regard the whole project as one that can be done in a ten year period.

You still do the first ten percent first.

CHAIRMAN BERG: Wally, in terms of these estimates, this -- There's a gap somewhere. We came to an estimate of what it costs to get the libraries, lambda libraries, cosmid libraries, YAC libraries. Then you're talking about what it costs to sequence, given that I have the M-13 clones already done. Somewhere there's a gap in between, which I haven't estimated -- gotten a feeling for what's -- I mean, Lee was talking about automating that whole process, and it's clear that the estimate depends on whether one can automate a lot of it. But since that's a missing number, can you give a feel for -- Is that a very large number relative to the sequencing? I mean, is it going to cost more than the sequencing?

MR. HOOD: Oh, I would guess it would cost on the order -- the same order of magnitude as the sequencing, or perhaps slightly less.

CHAIRMAN BERG: You mean, on a per base basis?

MR. GILBERT: No, look, to clone, prepare -- give an example.

MR. HOOD: I mean, we have to find where we're starting from.

MR. GILBERT: To do a megabase a year, that person who is doing a megabase a year has got to handle twenty-five cosmids a year. We have to have a technique available that gives them deletions, deletion patterns, for those 25 cosmids. That's one cosmid every two weeks or handling one cosmid every ten working days with some technique that's going to break that cosmid up into 100, roughly 100 subthoughts. That's not full time work for that person. It's a small fraction of the work that that person is doing. That's why I don't regard it in terms of the overall cost for the thing.

CHAIRMAN BERG: Well, let me put it this way then. Without the automation we still have a big job, and expensive in time.

MR. GILBERT: Without the automation, it's one person's effort.

CHAIRMAN BERG: To do 25 cosmids a year?

MR. GILBERT: Break up cosmids into 100 subclones and using a repetitive procedure in which they would make nested deletions.

CHAIRMAN BERG: Manually.

MR. GILBERT: Manually. And I give them two weeks to do it. One cosmid, two weeks to do it.

CHAIRMAN BERG: Well, you see, the added cost --

MR. GILBERT: And they're doing it over and over again. You've got one person's work.

CHAIRMAN BERG: Well, it depends on the lab, but we had a discussion once before. Depends on whose lab is doing it. But if you're saying one person per year could do 25 cosmids, breaking them up, getting them ready for sequencing --

MR. GILBERT: Just have to break them up into the subclones.

CHAIRMAN BERG: And get them ready for sequencing?

MR. GILBERT: Yeah. That's with breaking them up into --

CHAIRMAN BERG: So the most you would be saving by the automation is one man year?

MR. GILBERT: That's right.

CHAIRMAN BERG: Per megabase.

MR. WATSON: So who has done this so far? What is our ability that someone can do it?

MR. HOOD: No one has done it so far. These are -- I mean, and no one has done it, because there wasn't a need to do it before. What we did it on were smaller fragments that we could sequence directly from M-13, so you could carve those away and do the sequencing. Now we have the scale of things where we can think about jsut extending that up to, you know, one-quarter of magnitude larger fragments of DNA. But it hasn't -- I don't know, unless -- Have people at duPont played with this kind of thing at all, Mark?

MR. PEARSON: Well, we're playing with it, but I think you put the problem rather well. The dilemma at this point is that what we've done is move the burden of effort from running gels and reading sequence, I think back to the front end, preparing subs.

CHAIRMAN BERG: Seems to me, you might have a flight controller type problem, you know, people with battle fatigue, if all they're doing is just taking one megabase cosmids and then subcloning and making nexted --

MR. GILBERT: That's why you automate things, but the question -- In trying to estimate what the cost of something is, you have to start with --

CHAIRMAN BERG: Yeah, okay.

MR. GILBERT: -- you know, what is the amount of work that's involved, or some estimate for just how much physical labor is involved in doing something.

CHAIRMAN BERG: You're saying then, if the number we use is 150 k per megabase handling costs, isn't it?

MR. GILBERT: I'm saying that's an outside estimate.

CHAIRMAN BERG: And then, if you automate it, assuming that the machines don't cost you more than that, it will be less. Right?

MR. HOOD: Yes, a lot less. I mean, saying automated only improves -- only gives you one additional person isn't true in the sense that you'll be able to automate doing forty or eighty of these things simultaneously. So you're talking about enormous amplification capacity in that regard.

CHAIRMAN BERG: So then we just have to -- As an outside number we would have to multiply that number roughly 150 k for a man year times the number of megabases that have to be sequenced.

MR. GILBERT: For the handling.

CHAIRMAN BERG: For the handling. Yes. I just wanted to get an order of magnitude of that number. I think we've got an upper limit, and you're saying it could be less if it's automated.

MR. WATSON: You really think you have an upper limit, Lee?

MR. HOOD: Well, actually, if you want to put the upper limit on, I suppose you would have to really go to the random cloning and the need for sequencing ten t9imes over --

MR. WATSON: No. I mean, do we have enough experience that one could --

MR. HOOD: We haven't had enough experience. So I couldn't really -- I haven't -- I mean, what Wally says seems reasonable to me, but we haven't done it. Okay?

MR. PEARSON: If I could just comment further and amplify on some of the remarks Lee made with respect to the automation of sequencing per se. Number one, it can't be stressed too much that the technology development is moving rapidly. It's picking up in several different areas, as he's already indicated. And any major push on sequencing has got to have tied the two together that -- We've shown before that, if you have the technology guys and the people who want the answer separated, that's a mistake.

The issue that I would take, I think, with Lee's estimates and perhaps Wally's as well has to do with the bottom line number. I went through several scenarios of what it costs us in duPont to do sequencing now by conventional methods and what it looks like it's costing with the sequencer. Then I tried to take current costs and put a variance into each of the elements that I figure goes into doing a sequence.

What I come up with as our current cost is something like $1.25 with conventional film technology, and about twenty cents under our best estimate. And that's for one base pair determined once. So this doesn't have to do with the repetition factor, so that you start to feel very comfortable with the number.

So I take issue with the two cent to six cent number, but I'd be happy to see. Perhaps Lee could provide some numbers, and I can, to David who can go ahead and hassle through them.

I think the --

CHAIRMAN BERG: You're considerably higher than Wally's, because Wally's number was -- going through it several times in order to be sure you got a duplex sequence.

MR. PEARSON: Well, I wanted to point out that in my estimate I also went through our overhead to run a laboratory. So I'm assuming that we're paying for the laboratory and not just for the --

MR. HOOD: See, that's really the difference. If you bring in overhead, if you start amortizing instruments --

CHAIRMAN BERG: Somebody's got to pay that.

MR. HOOD: No, no. I agree; but I'm saying, we haven't taken into account those things, and they bring the cost up.

CHAIRMAN BERG: So it might be more like half a dollar a base pair, isn't it? I mean, once you start talking about using those numbers a number of times through, overhead --

MR. HOOD: Well, you're saying ten cents, twenty cents?

MR. PEARSON: I said twenty cents, yes. And just so you got a feel for it, I did it on a 300 base pair run. My assumption -- My total cost was 92 bucks, of which $50 was lab overhead, and this was after chopping back primers, enzymes, nucleotides, gels, assuming amortization of machines, both nucleotide syntheizers and DNA sequencers, over a three to five year period.

MR. HOOD: Sure. I would guess, with that --

CHAIRMAN BERG: Actual cost then would be roughly equivalent.

MR. PEARSON: I think we have to be very careful in feeding numbers forward to other people who aren't familiar with the technology to aprpeciate the real cost and not what it takes for your very best graduate student in their final year who is really cranking it through.

CHAIRMAN BERG: Right.

MR. COOK-DEEGAN: We took your figure. We multiplied by three, following the rationale that --

MR. PEARSON: Yes. I would point out, too, that there is an unknown in this game that, I take it, Lee and his people haven't solved yet, and neither have we. That is to develop smart software that is capable of pushing these error rates down. A ninety percent genetic map, if you will, a physical map, is very useful. A ninety percent accurate sequence is not very useful; and it's important for people to recognize that we need to get down to the kinds of below one percent error rate for sure, if that sequence is going to be more generally useful for clinical diagnostics, on the one hand, drug design, etc., on the other.

MR. BURKS: One quick question. I think this has been said in several ways indirectly. For this threefold sequencing or any other fold level that you can support, is there a general consensus on what error rate one should strive for? If you're talking --

MR. PEARSON: I think it's quite clear. What we'd like to do is sequence one strand once and know that that's it; but I don't think, based on reality and previous experience, anybody would feel comfortable with that now. I know that, from a machine perspective, we're trying to appreciate or have the machine appreciate some of the compressions and sequence anomalies that people in the field are well aware of.

MR. BURKS: I guess what I was looking for more is, given reality, what is an acceptable number? I mean, are biologists happy if you give them a sequence that's one in a thousand, one in a hundred error rate, one in ten thousand? Where do you feel you can cut off and say --

MR. PEARSON: Well, I suspect people -- If you could get down to one in 100 or slightly better than that --

MR. HOOD: I'd suggest if you could get down to one 1,000 or one in 10,000. That puts you in the ballpark of polymorphisms. I think people would probably feel comfortable with that.

The important point to make is, if you give experienced DNA sequencers the same gels and ask them to read them through for the first two or three hundred nucleotides, the error rate that even good people make can approach one percent. So, I mean, waht people don't appreciate is that probably the machines are beginning to approach what you can do manually now, and I think in a short period of time we're going to see what the manual analysis is.

CHAIRMAN BERG: One point which hasn't been raised, and I don't want to introduce it in the way of saying whose DNA are we going to sequence.

MR. HOOD: That's not a cost issue.

CHAIRMAN BERG: But the issue is, do you sequence just one, whichever one you choose? Are you satisfied with one? Is that the human genome, or in fact is the sequencing cost -- you have to incorporate some multiple individuals, however chosen --

MR. HOOD: Well, it seems to me, for the first effort what you want is nothing more than a roadmap to where human genes are. My guess would be the first one would be a composite of chromosomes from Japan and Europe and the U.S., and I don't think it matters a whit whether it's some linear combination as long as we get a roadmap that tells us where 99 percent of the 100-200,000 genes are.

Then the issues of polymorphism are quite different. There, you want to look at HLA or T-cell receptor and then really understand how the variation correlates with the disease and so forth. But I think in the first case, all you want is a roadmap. Whether you do ten pool people or whether you do three individuals and separate out the chromosomes and divide them up, I don't think it matters.

MR. COOK-DEEGAN You're saying sequence what's in your repositories, essentially, in your physical map repositories.

MR. PEARSON: In talking about sequencing strategies, there's one that hasn't been mentioned that at least ought to be flown up, and perhaps several of the people here, even in the audience, have thought about it. That has to do with amplification technology and direct sequencing without cloning.

It looks like that technology is useful, at least to a point; and I, for one, have no feel for the general limits of it. I would simply put it to this group that, if we had a finer structure map, Helen's 5 centimorgan map, for instance, we could in combination with some of the cloning techniques that have already been described -- we could in fact set out to start sequencing the human genome without getting the sequenced clones incorporated into appropriate vectors using conventional amplification technology.

CHAIRMAN BERG: Mark, that amplification technology requires that you know some sequences.

MR. PEARSON: I know that. I know that. But I'm saying that one could imagine -- One could conceptually imagine going ahead and doing a random sequencing with quasi-defined clones that are at least --

or quasi-defined primers that are derived from different regions of the genome. And I, for one --

MR. HOOD: One of the problems taht I think you would run into, Mark, is the polymorphism. When you clone, you get rid of the polymorphism anyway. When you don't, you would be looking at -- In the HLA region, it would be a mess to sort out the sequencing. I think that's a real concern.

Your ability to amplify is what? At the best, 2.5 kb or something like that. So it would be -- I think it would be an incredibly expensive endeavor, because you would have to make your primers for either random, and have them overlapping. So you would really want to look at the cost of --

MR. PEARSON: But remember, we're looking at big costs as it is to go through some of the cloning procedures, and I only throw it up because I've dismissed this for a long time, but I'm not sure it's so dismissable at this point in time. It might be worth thinking through. If others have ot thought about it, perhaps we should simply dispense with it.

MR. GILBERT: My own feeling about it is that it's a more expensive approach than the cosmid map and then its analysis, unless it changes radically to much longer pieces. Well, again a problem also about the accuracy of what you amplify and things like that.

I wanted to make a comment about accuracy in starting off and formulating this little project. There are costs involved in the quality control. There are costs involved in maintaining accuracy. There is some level of accuracy that needs to be specified as to what one would accept for a reference sequence, which I think my attitude is that it should be somewhere between one and 10^4 and one in 10^5 accuracy.

The source of inaccuracy, I think, is not the statistical inaccuracy of the programs or this one percent error, because that's caught most very easily by repeated readings, but it is features in the DNA which are unusually difficult to sequence; and those are rare and in fact it isn't even clear, I think, today how often such elements will arise in the sequence and what sort of level of problem they would finally achieve.

CHAIRMAN BERG: I'd like to raise a point which I'd like to get the sequencers to comment on.

One of the things that has struck me from the very beginning in thinking about this problem was that it's important that at every stage or as we move along in this project not to be waiting for nirvana so that we have the whole sequence, but to be able to get useful information along the way.

Technology development clearly spreads the benefits to lots of people. The genetic map, physical map, is useful to everybody long before we have a sequence. The question isn't on the sequencing project itself.

Seems to me it's going to be rather difficult to get people very excited about doing sequencing, even if it's done automated. I mean, all the tasks that will have to be done are essentially just grinding out something that doesn't seem to make much sense until we get it all together or in large hunks.

The question I'm asking is: Would it make sense as a strategy to essentially go for those parts of the genome which are interesting? Now how will I define interesting? I'll say it's expressed -- Let's say one were to in fact create as a first step cDNA libraries of a variety -- all human tissues, maybe at various stages of development, and then use those cDNAs as the probes to get the genes and do the sequencing on the genes.

We would get almost, seems to me like 70 or 80 percent of the sequence, because we'll incorporate not only the coding sequencing but the introns and some defined distance upstream and downstream. That makes sense, because we'll discover genes that people won't want to study. It will be interesting, rather than just doing blind seuencing.

MR. GILBERT: Yeah. What you say is perfectly true. However, what you haven't done -- and this is a cost meeting -- you didn't put a cost on that.

CHAIRMAN BERG: No, I -- That's why I want you to comment.

MR. GILBERT: But you actually should try to put a cost on that. I think the cost of that is several orders of magnitude larger than anything we've talked about yet.

CHAIRMAN BERG: cDNA cloning?

MR. GILBERT: Yes. Because the problem behind it -- Well, you do need full length clones --

CHAIRMAN BERG: No, you don't; not to use them as probes to get genes. To use them as probes, you only need bits and pieces.

MR. GILBERT: Sorry. And some of that depends on whether you think the genes, as spread out on the map, are reasonably dense in the DNA or are at great distances from each other, where you think the problem is.

I would argue, specifically, as to that, that if you want that idealized thing, the list of all possible genes which is what you think you're getting with the cDNA cloning, then in fact the problem becomes a biological one of a high order. It isn't sufficient just to take a few random cDNA clones. You want to know something like all of the minor cDNAs from all the tissues, and you'd like to have some measure about whether that's a complete set. It becomes a very difficult problem.

Doing the human sequence the other way around from the cosmids sequence, there's a lot of waste DNA in between but gives you all of the information, but not yet in the form in which you can -- you understand. You have to analyze that information. It gives you all forms of every gene, but you don't know how to put the exons together initially.

CHAIRMAN BERG: Well, I'm trying to introduce a little bit of interest into it.

MR. GILBERT: But I think the problem there again goes back to thinking what the project is. I say every time I'm asked, and since you've asked I will say again that I regard the underlying project of producing the human sequence as a production project, not suitable for the university laboratories, not even suitable for the National Laboratories, but in fact something which should be done as a production effort.

It takes a current scientific technique and will do it by production technicians, just like you were running clinical laboratories, just like the people making the drugs for the drug companies. And that's the group that will actually do it. It is not scientific research.

CHAIRMAN BERG: But that is -- I mean, this meeting is designed, obviously, not to meet that end. Why? Because I'm -- Well, it's not primarily. I assume we're talking about how the government can be advised to fund it, and I suppose you would say they could choose to fund it totally in the commercial sector, or --

MR. GILBERT: The government builds all kinds of things. The government can set up a production plant and make the sequence, as they want the government to do, you know, and provide the sequence to the scientific community so it can think about it.

MR. HOOD: I would see these two things as being concurrent. I think biologists are going to start from cDNAs or from genes that they -- where they have assays; and they'll go and sequence those things, and I think what will happen in the next few years is we'll be able to sequence, you know, 50 kilobases very easily.

CHAIRMAN BERG: One extreme to what --

MR. HOOD: But I think what will happen is some of the centers that are interested in sequencing, one or more could evolve into this production line center. I agree with Wally entirely. I think there you can make things incredibly streamlined and incredibly efficient, and the criteria isn't scientific interest. It's just getting the roadmap done.

I think this session is talking about both of those aspects of it, giving on the one hand biologists the tools to do things they think are interesting, yet creating a set of tools that will eventually give us the complete roadmap, and not piecemeal.

I agree with Wally again. I think it's incredibly cost inefficient to in bits and pieces pull out the cDNAs.

MR. GILBERT: The reason for having the sequence is to have essentially, I regard, a data base, a catalog of possible genes. That's where the scientific interest is. It's in the analysis of those genes; and from our view as scientists, we want the government to go give us this thing so we can work with it.

MR. COOK-DEEGAN: So assuming you're the government, when do you start funding sequencing efforts and at what level? Where clearly you've got a physical map?

MR. GILBERT: First of all, that's part of my answer to Paul's question is going to do that. I would sequence human DNA after having a physical map, so that I could sequence it chromosome by chromosome, and thus have coherent chunks of information coming out with considerable rapidity; because I think once you've sequenced a few chromosomes, you will then have a very good idea of the density information, how it lays with DNA, all of that sort of thing.

I would shoot for sequencing the DNA at a rate such that one sequences, as one is set up to finally dc it, sequences in the order of 10 percent of it a year. That gives you a coherent chunk of information.

In terms of sequencing strategy I would say let us set up sequencing production facilities. Let us have them sequence -- Let us have them demonstrate that they can sequence to a particular error criteria, speed and cost. Let us practice by sequencing a small eukaryote like yeast, and let us then go on the year after and sequence the first human chromosome.

The reason I argue that way is that I think the speed of sequencing from the cosmids, speed of sequencing from the physical map, is such -- I should say it this way. I think that the cheapness of sequencing the cosmids as a rote process is sufficient to make it worth sequencing the cosmids directly, as opposed to any strategy which tries to find just the most interesting part of the individual cosmids, because the strategy to look for the most interesting parts get caught up in an element that Lee mentioned earlier, the difficulty of finding the relevant DNA portions, the difficulty of subcloning and identifying just the right pieces to work with.

That's in fact where most of the effort in the current day sequencing goes.

CHAIRMAN BERG: With the cosmid -- Well, I don't want to belabor the point, but with the cosmids available, the cosmid libraries, obviously, cDNAs can be mapped to cosmids very quickly. Very quickly.

MR. GILBERT: But you have to have the cDNA.

CHAIRMAN BERG: Yeah, but that's trivial.

MR. GILBERT: No, it isn't trivial. Well -- Sorry.

CHAIRMAN BERG: You were going to say --? Well, never mind.

MR. WATSON: I'm glad you think it's trivial.

CHAIRMAN BERG: I think it's easier to get -- but what I said was to use them as probes.

MR. HOOD: Paul, I think you're right for maybe seventy or 80 percent of the genes. Okay? But let's go to the brain where -- what? maybe 60 or 70 percent of the messages that we think occur do occur, and many of them occur at the rate of one in a million or something like that, meaning they're represented in maybe just a few copies and a few cells. Those things, you wouldn't see, at least --

Maybe you could think about new strategies for cDNA cloning, but there is a whole repertoire of genes there that you would miss entirely by the kinds of things we've thought about now. I think that is an example of where going through and using computer programs to convert raw DNA sequence into 98 percent of the genes, -- that that's the predictability that exists now -- I think would be a very much more powerful strategy for getting that -- what is it? 15 or 20 or 25 percent that would be tough to come up with cDNAs.

The other 75 percent, you could get; but, you know, you would just be part of the way there. Why not do the whole roadmap and have everything? I mean, I think both are going to be done along the way, and I think people will do that.

CHAIRMAN BERG: Well, one is going to be done on one scale, and the other is going to be done at a two orders of magnitude higher scale. That's where the money will go.

MR. GILBERT: No. It depends on how the money is controlled and what the money is being used for. One of them is being done anyway. One of them is, let's go look for genes specifically.

CHAIRMAN BERG: No. People start looking for a gene of interest, and they work at that; but they're not in a systematic way. I mean, I could say that you could set up an operation which is specifically designed to pane cDNAs and their genomic counterpart, and that's the tool, the stuff for sequencing. The reason is that those were expressed.

MR. GILBERT: Think about the cost of it.

CHAIRMAN BERG: I'm glad you're educating me. I'm thinking more what I think from the point of view of what attracts people and why it holds their interest and what might return benefits long before the entire roadmap is finished.

MR. COOK-DEEGAN: Ruth, could you make a comment on the relationship between these -- I mean, it sounds like part of what the cDNA strategy is really the sort of thing that you guys already support in a big way.

MS. KIRSCHSTEIN: No, I don't think so. The only question --

MR. COOK-DEEGAN: In a sense that would plug into it, I should say.

MS. KIRSCHSTEIN: The only point I would like to ask Dr. Gilbert about is whether he believes that, if this is going to be done in a production, how you will assure that the error -- I mean, you can set up a pilot and assume that the error rate is going to be very low because these people have shown it, and you have to factor in what Paul was talking about, the fatigue, the boredom, etc., etc., ; and it has all the earmarks of the problems of good control on a contract.

MR. GILBERT: Absolutely.

MS. KIRSCHSTEIN: And that's very difficult.

MR. GILBERT: Maybe some multiplier should be added to account for quality control.

CHAIRMAN BERG: Frequent coffee breaks and vacations.

MS. KIRSCHSTEIN: It's a very difficult quality control problem.

MR. GILBERT: That's right, but you have the quality -- you know, in some ways we have a quality control problem anywhere in science. We get it by repetition, and the rest is science. We get it here by doing both strands.

MR. HOOD: I think that's a very important point, Ruth; because we're going to be sequencing two or three times over or four times over every piece of DNA, and you're going to have the software to play inconsistencies into evaluate it. I think having that feat is really first rate, is the key to the error control problem aspects. So you will have all of these internal checks going on, which is basically what error control is.

CHAIRMAN BERG: Christian?

MR. BURKS: I wanted to throw out sort of a perhaps rhetorical question about the evolution of these costs. A couple of years ago when people were first talking about the question of mapping and/or sequencing the genome, the standard cost then people tossed around then was a dollar per base; and everyone's response was, that's just too expensive to envision going after the complete genome, let's do technology development and then talk about sequencing.

Well, it's now two years later, and we're talking about ten to twenty cents a base; and my question is, even if someone said let's do the factory approach, it would probably be a couple of years from now before that factory was actually built and installed and hired people and was working.

Would one expect with the other -- Mark alluded to possible very different technologies. I know there's a group at Los Alamos that's interested in evaluating entirely different technology. Is there any thought that two years from now that we'll be down to one cent a base or something significantly below the ten cents a base domain for how much it would cost?

I raise the question. I don't have the answer to that.

MR. HOOD: Well, I guess my feeling is, if you wanted to take these technologies and improve them to their ultimate and then design instruments for these centers that were, you know, five or ten times as big as what an ordinary lab would use, I think you could easily get a cost factor of five over a period of two or three or four years.

So my own feeling is you could cut it by a factor of five without doing any kind of new technology whatsoever, just optimization. If other approaches start to work, I mean, that's a different thing.

MR. PEARSON: I think it's very important to realize that the current technology and sort of the near term stuff has all been developed with our previous experience in mind, and we're kind of designing technology for individual operators in a standard molecular biology lab. To really develop megasequencing capabilities, as Lee is indicating, I think, sure, we could push down the cost; but realize, too, that int he end, if it costs nothing to do all the sequencing, to be facetious about it, still half the cost is the damned overhead to have a lab in which to do nothing.

I'm only trying to point out that this minimal cost, in the end, can only be driven down by finding a way of doing it in a very cheap, low overhead environment, and that's one of the points that I think Wally is driving at.

MR. WATSON: Even with -- nested deletions, I think we would be irresponsible to claim we will ever get it down to less than ten cents a base pair. Or do you think you can get it down to one?

MR. HOOD: Well, I think we can probably get it down to less than ten cents a base pair. You know, it just really depends on the scaling factor, and it depends how much the overhead and the amortization and all of these things slow down that whole rate.

So I guess I don't know, but I think you can get it under ten cents a base.

MR. WATSON: But should we be worried about ten cents a base pair? I mean, that's only $600 million over probably fifteen years. So it's not that hard to sell.

MR. GILBERT: I think you're absolutely right, because if you spent $300 million to lower it, you wouldn't be anywhere. It's only if you can lower it by easy things -- I think it's at that level at a range in which the project is reasonable to do.

CHAIRMAN BERG: But I don't think we should be in a position of promising it's cheaper to reduce the cost beyond which we don't really have to reduce it. Well, having to -- it depends on how optimistic you are that somebody is going to give you the money in the first place. I mean, what do you mean, if you are going to ask for 3 billion and you can get it down to 600 million or you can get it down to 200 million, that makes a big difference, coesn't it?

MR. WATSON: But it's 600 million. And you say that's over fifteen years. That's $45 million. It's one percent of the NIH budget. I mean, it is not a colossal relative amount of money.

CHAIRMAN BERG: Okay. Bob.

MR. STEVENSON: Can I ask you -- Not being involved with any of this, I can ask stupid questions and get away with it. What is the final product that is envisioned? Is it going to be a thick volume with a bunch of sequences, unindexed with nothing to represent the sequences? Is it going to be a fully indexed thing with meaningful sequences represented by physical materials in a bank someplace? What is it exactly that you guys want?

CHAIRMAN BERG: That very silly question is going to give everybody pause. So we're going to break for five minutes while everybody thinks about it, and then we'll come back and try and respond.

(WHEREUPON, the workshop recessed briefly at 2:31 and resumed at 2:55 p.m.)

CHAIRMAN BERG: To put us on track to where we stopped a little while ago, we've come to the point where, presumably, we will have a genetic map, a physical map, a library of overlapping cosmids. We've discussed what it will take to take the bits of DNA that have been cloned, convert them into a sequencable form, and the actual sequencing cost. As Wally says, in the end what we all want is that string of letters which defines the entire sequence of the human genome.

I think the question that was posed at the very end was what are we going to do with it, and how are we going to use it, and how are we going to evaluate or interpret all the extraordinary information which we're counting on is buried in that string of letters.

That has to be costed in, because that's not going to be a trivial cost. So, Christian, I gather you've thought a lot about this. So I'm going to ask you to perhaps address -- I gather you deal with the sequences that are available today and the problems that are associated with that. And if we now go up by a factor of 10^3 or something, or 10^6 over what we have, how are we going to deal with that?

MR. BURKS: Okay. Yeah, I have put some thought into it and have some comments. I was getting nervous earlier whenever the word computer arose, the word trivial cost was quoted. (laughter) It's certainly not our experience in dealing with GenBank and the other projects we're interested in, and I don't think the NIH would agree with that either at this point.

I guess -- sort of introductory comments before I get into some of the dollar estimates that I came up with. One is that, over the -- I guess it was close to a year and a half ago after the Santa Fe workshop when -- in talking with some of the people who participated there, we came up with a budget projection which was included in this pamphlet here.

In the computer end we talked predominantly about a very centralized super-computer or mainframe computer resource that would serve a project like this. i think -- Certainly, my thinking and the thinking of people I work with over the last couple of years has, though not abandoning that as perhaps an important asset and resource on a project like this, has come around to the notion of emphasizing a great deal of distributed processing, especially to the extent that there isn't a single building that is doing mapping, sequencing analysis and associated science.Where you have different small groups of people around the country or around the world working on this, the technology and tools are now available to support them very much locally in terms of their computational needs to manage the data and to analyze the data.

So there are sort of three modules that I'm going to describe here. The first module is basically, I'd say, an R&D group that sits down and thinks about managing all this data and linking different data sets to other data sets and organizing the data in a way that allows people to maintain it and query it in as efficient a way as possible.

The second module I'll describe is basically a sort of functional mini-computer center for any of these, say, big labs that are going to be going after sequence or mapping data in a relatively directed way.

Then, finally, I'll get to the thing we described a couple of years ago, which was the sort of central computer facility. The reason I'm doing it this way is it's very hard for me to -- and I think anyone, at this point, to estimate, for instance, how many labs will be really tied into this network of information gathering and use.

What I have attempted to do is sort of describe the costs per lab, and leave it to further deliberations on how many units there are like that will be existing and tied together to come up with a sort of total cost for any of these small modules that are stacked up.

So first, for an R&D aspect -- I should also preface by saying I used the 100 k per person year value, and that can be replaced with 150 in a retrospective viewing of any of these numbers, if that seems undesirable. I'm all for it, of course.

The R&D -- The module I had in mind here was something on the order of six staff, and in terms of time lines, this seems to me like, obviously, the first thing that should get going. I mean, there's nothing at this point before there is any overwhelming amount of data to prevent -- in fact, it should be encouraged -- one or more groups like this to get started immediately on anticipating the data that's going to be generated, how much there will be, what kind of links one will want to develop between the data, how one will want to network the various data gathering stations, and whether they be sequenator based or a more traditional molecular biology group working on the mapping data, anything in that broad spectrum of things.

So six staff at 100 k a piece, I come up with 600 k per year, and then, of course, they need a hardware and software environment to play with in coming up with applications of whatever thought they have about how to organize and manage the data. The number I have there is 260 k per year. In my next module I'll describe in later detail how I came up with 260 k.

At any rate, a R&D module, a gorup of people thinking about how to organize the data, I come up with 860 k per year. Of course, there could be more than one of those. I mean, there's a group like that that started up at Los Alamos with support from the DOE. I'm sure there are groups or groups that wouldn't identify themselves as such that are doing work like that under Howard Hughes and NIH through other projects.

Now I'll leave that for the time being. That's one module at 860 k per year. The second thing I wanted to get to was this local mini-computer center environment that an R&D group could depend on, but also that any of these large lab or distributing data gathering centers could rely on for their work.

What I've done there is basically come up with a list of hardware and software for a personal work station environment for a small group of people with enough staff to maintain and manage that system. On my shopping list there includes a sort of upper end server work station, two lower end DISCIS and other work stations, two large scale storage devices, an IBM-PCAT or something like that that would connect you to the rest of the world. I didn't have access to what might be considered the more pioneering hardware that's available. A laser high speed printer, and supplies, tapes, floppies and so on, networking hardware to connect you with the other nodes that one wants to be connected with, and an annual maintenance fee. That all would be a sort of first year outlay of 100 k.

If you looked at an annual increment based on the maintenance contracts, supplies and increasing your memory capacity, disk storage capacity, that would come out at about 21 k per year. I throw that into a four year cycle; that is, that one would -- That's a little bit of a soft number. One would say that every four years you repeat that. You buy all new equipment, because you can do much better either for the same price or for cheaper prices, get the same kind of performance.

So if you average those numbers and I also want to throw in numbers for software and staffing, you come up with the 260 k per year. I'll get back to the 260 k per year in a second.

I've just covered the hardware part of it. The software part is dominated by, as I budgeted it in here, a purchase of a commercial DBMS, relational DBMS system, that would be available on all of the work stations included int he shopping list; and that, plus maintenance, again a one time outlay the first year would be something like 49 k, and the increment on that would be about 9 k per year.

Finally, with staffing I came up with two FTEs each year with responsibilities including system manager for the hardware, DBMS manager for the data base management system, and an applications programmer to do custom software for whatever is going on in that particular environment. At 100 k for each of those two programmer/managers, I come up with 200 k there.

So again, averaging over a four year cycle, summing up the software staffing and hardware numbers there, I come up with 260 k per year for a local work station environment for any particular lab that's generating data. That was the number that I plugged in, talking about that R&D group earlier, that 260 k per year.

The final thing which these numbers are a lot softer from my point of view perhaps because I'm not expert from a hands on point of view in managing or acquiring a super computer, but I've relied on contacts with people at Los Alamos and elsewhere to provide me with some of these numbers. One can certainly imagine a centralized computing facility; certainly, in the case of some of the kinds of analyses, not necessarily end analyses but in process analyses, for stitching some of these kinds of data together, using currently algorithms -- current algorithms, if you were doing -- In some part of your hierarchy of how you go about the sequence collection effort, if you were doing random sequencing and using a computer to stitch together fragments, that can very rapidly turn into a very CPU intensive task, even with some of the advanced, high speed, and hashing out rhythms and other approaches that people have developed.

So there are certainly tasks that one can imagine using a much larger computer. I don't know whether a super computer is what would be required, but I've gone to the other end of the spectrum here in this shopping list and imagined something on the order of a CRAY.

A CRAY costs about 15 million to purchase, about 3 million -- I think Jake Musel earlier gave me the number 20 million. So my numbers might be old-fashioned there. Fifteen million to purchase, 3 million per year including support staff to maintain. I came up with a somewhat imaginary number for peripheral hardware of a million, and maintenance on that about 150 k per year.

Then looking at the maintenance contracts on the CRAY and the peripheral hardware, that would come up to an annual increment of $3,150,000 per year. On the software, that's also in the area of the CRAYs. I'm not as cognizant of what kinds of costs would be involved there. I threw in a number of 2 million in the first year for software, perhaps DBMSs and other things that one has to go out and buy that don't come with the machine or the operating system.

Upgrades and maintenance would run about fifteen percentof the typical number. That would be 300 k, and one can imagine annual increments of 300 k there.

Finally, staffing -- If one looks at something on the order of the GenBank project or some of the other projects where people have a staff that's clustered around the CRAY and are trying to do -- not only maintain a CRAY but provide support for people who need data, a good guess number, I think, is about 30 TEs; and at 100 k each, that comes up to $3 million per year.

Looking at that as -- Oh, and I would also throw in, of course, that locally sort of linked to the CRAY and the mainframe is you would want to have some of these work station environments that I described in the separate labs earlier. Again, I threw in a number there of seven. That would put one work station monitor in front of each of the thirty people that I attached to the CRAY in that staff. At 260 k each, that comes up to 1,820,000.

Averaging all those over an eight year cycle, again in a work station environment I talked about a four year cycle for purchase of new equipment -- I'm guessing again a little bit, but imagine an eight year cycle before you go out and buy a brand new CRAY. Sounds like Ruth thinks it should be a longer number. At any rate, averaging that initial outlay for the CRAY and associated materials, and then the annual increment for maintenance over an eight year cycle, I came up with 10,520,000 per year.

So I've sort of got three modules. I'll retrack here: The one I just covered for a super computer central facility, cost of about 10,520,000 per year; a work station environment that one would expect at perhaps even relatively small laboratories that are generating data that came up to 260 k per year; and research and development staff who spend time thinking about how these various stations should be tied together, what the best way to structure the data is and so on, coming up at 860,000 per year.

CHAIRMAN BERG: Do I understand that it's in the R&D module that you have in mind that somebody is going to be thinking about how to use this information? I mean, what are the kinds of search programs, the various -- I mean, this is a different order of magnitude than anything that anybody has done before. I assume we're going to need some wholly new methodolgy to be able to use this information.

MR. BURKS: In fact, that is a large part of what the R&D people would be doing. But I'd say also that a lot of the -- It's really split. A lot of the strategy for being able to answer questions about the data are built into the DBMSs, the commercial DBMSs that

I -- in this scenario -- described buying and plugging in.

So you don't have to develop that. You have to figure out perhaps how to install your data in the system so that you can apply the sophistication of the DBMS to your data, and the R&D people would certainly be working on that. But you're buying the software. Someone already has done a lot of work in a very general sense in setting up the software system that allows you to do very sophisticated queries and to maintain the data in as efficient way as possible.

CHAIRMAN BERG: Is current search technology capable of handling sequences of this size?

MR. BURKS: The technology is capable of handling sequences of this size. The question is -- In other words, the length of a sequence isn't a problem if -- I don't know what kind of queries you have in mind. If you're saying compare one sequence against the data base, it really doesn't matter how long the query or the data base sequences are to simply do that. Where there's a big impact is on how long it's going to take to do it on any particular machine.

I think there's a lot of work that is going on now and will continue to go on on algorithms that allow you to decide for yourself what tradeoffs you want to make. Do you want a very quick search that perhaps misses some things you might have been interested in just because you need a quick answer, or do you want to chug away for hours on end to get a very exact list of the things that you've scored against.

Lee, I'd be inaterested in any comments. You mentioned earlier that you had been thinking about some global picture, and I know that your group has been thinking actively about sequenator speeding into the data bases.

MR. HOOD: Well, we're going to do two things. One: Tim Hunkapiller has been playing with the systolic chips that TRW has developed which have the capacity, under specialized search conditions, to speed up by orders of magnitude data searches.

So, for example, you can build into these chips a whole series of conditional questions such as the nature of regulatory sequences, and then zip through the data bank very, very quickly and pick out just exactly what you're looking for.

So that's one application we think can be developed to where you can put it on, in the future, PCs and really extend their analytic power for certain kinds of pattern analysis.

The second thing we're going to do, together with the computer scientists at Cal Tech and JPL, is compare in detail the NQ with these connected machines with the CRAY on a whole series of kind of classic problems set up for data analysis and larage data banks and see --

MR. BURKS: So it's using multiple processors.

MR. HOOD: It's using multiple processors and getting at the whole question of what's the most efficient way of doing searches. Then that gets you into thinking about how you set up data bases.

MR. BURKS: I think that's an important point. I was stressing software, and I think hardware will also be an area where people are trying to cut corners on the amount of time it takes.

CHAIRMAN BERG: Ben, did you want to --?

VOICE FROM THE AUDIENCE: I appreciate that your model, to some extent, is accurate and, of course, scaling up. The thing we know how to do best, which is sequence comparison, very low level, determining the similarity of one thing to another thing in the data base; but I would submit that there's a whole higher level of questions that are going to have to accompany your genome project.

Those are software that allow the biologists to state a biological question and have all of the manipulations, mathematical manipulations, to test that hytpothesis be generated by the software, applied to the data base and as well, as you mentioned, the ability to detect possible areas of interest. I mean, what are the possible genes, but take it farther. Since I only have limited laboratory resources and a limited number of gels, what are the areas of -- the regions that are most likely to return the positive sequence that I should actually put in the test tube.

For that, you're talking about not just low level similarity determinations, but being able to encode very complicated huristics and very complicated rule based systems what comes under the general rubric of artificial intelligence, things that we have only the vaguest notion of how to do and which are very reasonably handled in these days, for example, by grant programs.

I didn't hear you mention that, that this has to have not just a small R&D group but probably a fairly substantial program that represents -- that allows a number of different investigators to pursue lines of development and research on a national scale to support this future level of software.

MR. BURKS: But what I -- I mean, tell me if I, in fact, am not addressing your point. My R&D group of six people there was referred to as a module by me, meaning that I really think of a group working on one aspect of any of these problems. I'm not claiming that I think six people is sufficient. In fact, I claime quite otherwise, that you would want several or many of these modules; that is, small groups of people working on various aspects of the question of how do we want to ask questions of these various data sets and how will we be able to get the answers out to answer those questions.

So I think I agree with you. I was really looking at a building block. I can imagine many of those blocks, somewhat at the discretion of the needs of the project and at the discretion of funding agencies being used to answer many different questions along those lines.

MR. COOK-DEEGAN: Ben, you said a substantial granting program. Could you quantitate substantial. Do you have an idea of what sort -- your annual budget would be for that sort of grant program?

VOICE FROM THE AUDIENCE: Well, you know, one way would be to -- It's possible that in fact a university based R&D program. I think for that perhaps the cost may be a little high, but even if we take that -- Let's say it is 750,000 as an R01 grant, allowing having the resources to support perhaps eight to ten of those per year, would allow you reasonably good opportunity of generating enough new ideas that you could have the software keep pace with the data.

MR. HOOD: You know, the one thing I would emphasize, though, it is really important for those groups to have connections with the real molecular biologists and biologists doing it, because if they are disengaged you get wonderful models that don't bear on reality, necessarily. But if you put them together, then they can cross-fertilize one another.

So I think having these in the universities is a very powerful place to generate these new kinds of ideas that we're going to need for this kind of thing.

VOICE FROM THE AUDIENCE: The two other components of research and development which have not been recognized, I think, or at least ought to get some focus in this -- The first of those is better user interfaces, because we clearly are not going to make molecular biologists into computer scientists. There just are not enough years in their lives or interest on their part to do that. So computers have got to get much, much more capable in terms of interpreting the vocabularies and the ideas that they express to their colleagues, so that the computer can be like an intelligent lab assistant. That's not developed yet.

The other is that there is as well this obligation for training. specifically, programs will need to be developed to bring biologists into the sphere of using the computer as a tool much as he uses gels and other sorts of things.

MR. BURKS: Ben, did I hear you say you

thought six to eight of these R&D groups would be --?

VOICE FROM THE AUDIENCE: Well, I don't know the optimal number. I think that you can probably look at other areas where we're looking -- we're trying to fund the development of new concepts of entirely new paradigms of analyzing complex areas, and perhaps use those as models for deciding what the optimal number is. It's certainly not just two or three.

MR. BURKS: That sort of leads me into, I guess, the questions I wanted to pose for discussion at the end of all this. One question is, of course, how many of these R&D modules do you want to have? How soon? How many? -- in order to begin to address the data. The second is how many labs do we think will be wanting to be tied together in order to produce this data in a way that it's tied closely together, that people in one place can know about progress in another place; that kind of thing.

That, of course -- All I've done is give building block numbers and come up with total costs which is the object of this exercise. One needs to know how many building blocks one wants.

The final thing, of course, is does one need a super computer center? Is that a part of --?

MR. HOOD: You know, to sharpen that question, I think a critical issue is whether every automated sequencer in the future ought not to have formatting programs so that that could dump their data directly into the data banks. I mean, the trouble that GenBank has right now is it takes forever to get the data in, and it's often wrong, and it isn't annotated correctly, and on and on.

You could build in programs, every computer that formats it, just as soon as you got the finished data, and then takes it on line to wherever the central -- And I think that's how you answer the problem of communication.

Now when people are willing to release their sequences and how we're going to publish sequences in the future is quite a different issue. I think -- But I would really envirion in the future we would have that kind of direct interface, and there would be one or a few centers that would be the repository.

MR. BURKS: Well,I've very much, I think, assumed what you've just said in the sense that I didn't include data entry staff in any of those calculations. I assumed that -- especially assumed some sort of cohesive approach to a project like this, that it will be more or less a requirement that those generating the data are doing it in such a way that it streams into these work stations or even a central facility in a completely or almost completely automatic fashion.

VOICE FROM THE AUDIENCE: But that's not a trivial aspect of it, I must say.

MR. BURKS: Yeah. I guess what I'm saying is, I think that goes under the sequenator cost. The data entry part of this, I think, goes under the sequenator cost more than into the computer support cost, though there may well, of course, be research aspects of that; but I'm throwing it back on the sources of data generation.

MR. COOK-DEEGAN: Getting back to that question again, does anybody have any idea how many of these modules do we need of the 260 k a year amount?

MR. HOOD: That was 750, wasn't it?

MR. BURKS: That was the big one.

MR. COOK-DEEGAN: That's the R&D groups. This is kind of the output for whatever groups are doing input. Right?

MR. BURKS: Right. There was 260 k per year for a computer environment that serves a data collection effort, and the 860 per year was for these R&D groups that are thinking about how this whole system should be set up.

MR. HOOD: How would you define a data collection group? Is that a group that produces more than X nucleotide per unit at a time?

MR. BURKS: I think what I have in mind is almost sort of more administrative definition. That is, one of those groups that's been invited to participate in this directed data collection effort and -- I mean, if for whatever reason they're only generating 1,000 nucleotides a week or a month or whatever, you know, they still aren't -- They wouldn't have been invited and be part of this unless there was something important about what they were doing, and that they should be tied into the computer network also.

MR. HOOD: See, it gets very complicated, because I think Ruth would argue, they're supporting lots and lots of genomic sequencing now. So the boundary of who is invited, who isn't invited -- I mean, I'm just not clear exactly what that unit is.

MR. GILBERT: Just to simplify, sort of a de minimus consideration. Any group that's sequencing fifty million bases a year can have one of these.

MR. HOOD: I think that's the kind of definition you want in some way.

CHAIRMAN BERG: George?

MR. CAHILL: Well, just at a much simpler level, the New Haven gene mapping data base or bases, five there -- in other words, five different data bases-- and tying that in with Victor McKusick's OMEM dealing with inheritance of man which the National Library put into a data base, and we're linking that to New Haven and tying that in with the mouse data base at Jackson Bar Harbor looking for homologies. We're putting $2 million a year into keeping those relatively simpler data bases going, to give you an idea; and we're talking magnitudes greater complication when you're getting into primary sequencing and searching and data accessing.

That includes -- That 2 million includes eight or nine editorial notes which is simply a PC and a part-time programmer. Toronto, Tom Shows at Roswell Park, Phyllis McAlpine in Manitoba, etc., to edit this RFLP plus DNA probes plus contact plus a literature data base to the five data bases in New Haven.

This is just to give you an idea. It's not cheap. IT's highly complicated, and data access is only as good as the data put in, meaning editorial screening. That's where the expenses are. That's manpower. There is no other way to do it, though there are problems. I can see the complexity of what you are trying to get into, is to put in the generated DNA sequences, be sure they're in there correctly from multiple sources. That's going to be an immense problem.

MR. HOOD: Well, see, what I'm saying, George, we probably can handle that in part by adding onto to the sequences and formatting what you need for doing that --

MR. CAHILL: But how are you going to do the editing, Lee? The question I have is the editing. In other words, the guy that turns a knob when he turns on a sequenator, are you going to be out of phase? In other words, you're going to have to have built into it monitoring.

MR. BURKS: That's the challenge to the R&D people, is to figure out how to make sure that all these data sets are not redundant and incorrect with respect to the same data set at another site, and that kind of thing. I think the data base -- The current data bases, I think, are in a phase where that they're using a model for data base work that was appropriate, say, when they came to be five years ago or ten years ago or whenever they came to be, but are being driven by the sheer volume of data to consider an alternative model and the use of tools that have come into being in the last five or ten years. We will again push, as Lee was describing, push the actual data collection and annotation effort off of the central computer staff onto the shoulders of the machines or people who are generating the data.

MR. HOOD: There are really two things you're worried about. One is the error rate, and you have to -- We talked about ways we can build in checks for that.

The other is the annotation. That really is straightforward. You can format that and, before you dump it, you have to have that all correct. The investigator can fill that out, and so forth.

So I think your problems could be significantly less than the more descriptive data where it is harder to know how to describe many different kinds of things.

CHAIRMAN BERG: Wally, can I ask you to explore your definition of a place that's doing fifty million base pairs a year. How many such places do you imagine there will be?

MR. GILBERT: Well, I used that figure because that comes out to six. If you think the goal is 100 million bases a year, which is a way of doing the whole thing in ten years, then you can cut the pie any number of ways. Whether it's a pie you want to cut this way is another issue.

CHAIRMAN BERG: I just -- That's your rationale for the fifty --

MR. COOK-DEEGAN: If you use the criterion of thirty megabases a year, then you need ten of them.

MR. GILBERT: It's a sort of way of saying it.

I think if you try to break it up -- Again, it's a question of -- To the extent that it is a repetitive production problem, there is a minimum size that a group has to be in order to function; and beyond that it's just iteration of whatever that size is.

CHAIRMAN BERG: Now is what Christian described --

MR. GILBERT: There's a research element in this that I think Christian was referring to, because there's a sense in which -- and it sums up on some fundamental level the goal of this is to create a data base that is a collection of things on a computer, not in a book, which has both been analyzed as to the best extend that our current machine technology can analyze it for us, and is available to the user to ask questions about, where the user has available again the best programs available at any one time to ask questions; and the questions are both in the form of what is there, what's related to what I have, what's known about what's there, how can I see it all, what is the way that enables me to look at any piece of the information that seems to me to be relevant.

MR. BURKS: I'm sort of throwing the coin at that six centers for the sequencing. There are other data sets that are being -- I mean, there are repositories of experimental information that, you know, one would -- I think you would consider those as nodes in this network. There are generators of mapping information. I think you would consider them nodes in this, too.

So I just don't want to focus too much on the number six right now.

MR. GILBERT: Well, it could be -- I mean, you can think of that as all aone centralized data base.

MR. BURKS: Logically, it's one single data base. As to whether it's physically centralized or not depends on the implementation and the direction you decide to move in after doing research on the best way to set it up.

CHAIRMAN BERG: So does the whole price package cost? I mean, we're supposed to come up with some kind of an estimate. There isn't anybody outside this group that's going to come up with a better number of how many such centers do you need, how many such R&D groups do you need, how many such super computers do you need, and how many grants of the type we were talking about does one need?

We've got unit costs. Right? So the question is: What does the whole package come to in a reasonable estimate, with error bars?

VOICE FROM THE AUDIENCE: Well, it might be reasonable to think about it as a percentage of the overall investment in the research. We did this exercise that is now about a year out of date, and approximately fifteen percent of the amount of the total investment in the research could reasonably and fruitfully be devoted to the information acquisition, analysis, and distribution of the data that we received.

I don't have the line by line derivation of that percentage, but we can show it to you.

MR. HOOD: I think various of us have guessed at this, and between ten and fifteen percent is usually the number you end up coming up with if you add up everything and say how much you'd like to go into the computing end of it.

MR. GILBERT: There is one -- There's two aspects to that computing. One is the holding and presentation of the information, the data base. The other is the analysis you were referring to before for deeper thought that's being put into it, and the amount of money you may want to spend on the analysis will be higher than --

CHAIRMAN BERG: That's open-ended, too. That's actually open-ended.

MR. GILBERT: It is open-ended, because it in fact refers to --

CHAIRMAN BERG: But they have different curves, don't they? I mean, you would obviously need to develop the storage capability earlier, and the analysis essentially, as the data becomes more and more complex. So as a rule, does anybody here have a better way of casting that number, roughly fifteen percent of what the total will be for all the other things, just as a -- Does that help you?

MR. WATSON: Could I ask whether your Los alamos numbers have any -- I mean, we were given between a half billion to 2.5 billion. That's much more than fifteen percent.

MR. BURKS: Say that again. I'm sorry.

CHAIRMAN BERG: Jim, can you repeat that?

MR. WATSON: The figures in the Los Alamos --

MR. BURKS: -- were a half billion to 2.5. That was for the entire project. Computing was just one component of that calculation. If you're referring to that paper that's included there --

MR. WATSON: I thought this was -- I saw it in the wrong way.

MR. BURKS: I think it was -- It was just under a million per year for the computing costs, I think, in that particular estimate.

MR. WATSON: Just under a million?

MR. CAHILL: That was just primary sequence storage, though, probably.

MR. WATSON: It's fairly high.

MR. BURKS: There's one section there on computing.

CHAIRMAN BERG: I want to, in the time remaining, because I think we don't want to skip it since that also seems to be one of the things that most people who are not wild about this project seem to focus on as a reason why we shouldn't be doing it -- So if we can just wind up the computing estimate, I want to get on to this last item which we are supposed to have started at 1:45. Okay?

So Christian, if you can -- you and Jim -- Do we want to use this number, roughly, that 15 percent or the best --?

MR. BURKS: Well, I think -- I mean, what I would like to do is to -- after this meeting when we know what it's fifteen percent of, which I don't think we know yet, it will be interesting to compare that.

CHAIRMAN BERG: But your number is flexible, because you left the n out. I mean, you have a unit cost and times n, and we don't know what n is. The n is purely judgmental at the moment.

MR. BURKS: This last session that's more on organization is going to supply a lot of the n, I would guess.

CHAIRMAN BERG: Really?

MR. GILBERT: The high price, in your estimate, is the CRAY.

MR. BURKS: Yeah.

MR. GILBERT: Now is there a reason to have the

CRAY as opposed to other approaches to the analysis and complex data?

CHAIRMAN BERG: Do you want to respond to that?

MR. BURKS: Jake Mazell maybe has something to say about that.

MR. MAZELL: I would just interject that don't be overawed by the CRAY. It's -- Of any super computer it is likely to be ten to 100 times faster than the next best super -- next best computer around. As we get into the point where we can predict a structure, we're going to blow lots of super computer equivalents to get the data that we're all going to want out of this; and it's no problem to blow away a super computer with this data.

CHAIRMAN BERG: What do you mean by blow away?

MR. MAZELL: Keep it busy. Not just calculating --

MR. GILBERT: We have to be careful with the protein structure problem. So far there is nothing in the DNA problem that is comparable to that.

MR. MAZELL: True.

MR. GILBERT: So at one level in terms of the analysis that we immediately foresee for the data base, I would have thought something, -- As I say, the connection machine, something a lot cheaper than the CRAY at $10 million a year running costs will do a great deal for the center of this project.

There's a research element to the project which, yes, I think will respond to the super computing centers and things like that. If I were designing it, I would look along those lines.

MR. BURKS: I think Jake's point is well taken in that, if you're really talking about, say, a ten year time limit -- I mean, I don't think you want to focus in too much on, say, one single kind of thing that we think we want to do with the data right now. I mean, it is now routinely desired, though most people don't have the computing power to do it, to compare every new sequence against the data base.

So, certainly, things like multiple processors or better algorithms lead one to imagine you could do something like that with GenBank today with something less than a CRAY.

MR. GILBERT: Well, certainly, you can with GenBank; but we're talking about something a lot --

MR. BURKS: What I'm saying is, with a much larger data base and with the result in the next ten years of the other kinds of things people will want routinely done and attached to the sequences that they get out of this project, you know, I think it's very easy to envision a much larager CPU requirement for attaching things like that.

For instance, if you're turning out a megabase a day and you want to give experimentalists a quick view of where the potential protein coding ranges are, that -- that may turn out to be a very trivial calculation; it may not be a trivial calculation.

There may be a hundred things like that that you'd like to do for every sequence that you get out of your sequenators. So I think -- I'm not presenting this necessarily that I'm saying we need a CRAY, but I think it's wrong to say that, looking at what we do with sequences today and ignoring all the other kinds of data sets that are going to be coming out of this, that we -- no, we don't need a CRAY.

MR. HOOD: But I guess the other point you could make is computational power is changing enormously.

And again, with these chips I think you can envision really cheap fast ways of doing the most computationally intensive calculations that we have to, dealing with primary sequencing.

MR. BURKS: Computing is always getting cheaper. I guess you either say I want the same power for less dollars next year, or for the same dollars I'm going to get more power. I mean, I don't know what direction necessarily --

CHAIRMAN BERG: Jake?

MR. MAZELL: I think in the very near future that we're talking about here, the super computer still has the great advantage that it's jsut like a big -- It's not like trying to design the algorithm at the byte level into a chip which you've got then and you're stuck with. You can't really deal with it.

Now eventually I'm sure that the individual chips, as they do now, outperform super computers for a specific task. Graphics now is done on a PC at a rate at super computer speed. So, you know, in certain limited ways it doesn't let you calculate the energy of the atomic -- There's a lot that you still don't know how to build into chips that the super computer is very good at.

CHAIRMAN BERG: Last comment on this.

VOICE FROM THE AUDIENCE: I would say that the emphasis on super computing and analysis is probably misfocused. As the data base grows, the major costs of the data become in the maintenance of the resources, that is, gathering it in, doing quality control and distributing it to an ever large number of sites.

Because it's volatile data that changes very rapidly, you're having to make multiple copies or multiple means of distribution. And so it's the data bank costs that begin to scale in a very large measure linearly or close to linearly with the data accrual. The analysis remains a relatively constant dollar amount, and for which you buy a bigger, faster computer.

Our experience with large scale data banks like Medline and these other molecular biology data banks that the NIH supports shows that it is largely the personnel costs, maintaining the integrity of the data, and the simple media costs, the distribution costs of getting it to the people who need it that really become the lion's share of the dollars devoted to the data.

MR. COOK-DEEGAN: We haven't plugged that in at all.

MR. GILBERT: Do you have an estimate of that?

VOICE FROM THE AUDIENCE: Not off the top of my head, but --

MR. COOK-DEEGAN: We would appreciate something like that. We haven't even plugged that in at all.

MR. GILBERT: In a sense this data base is much smaller than --

VOICE FROM THE AUDIENCE: It's about 1/20th the size of Medline. We do know those costs, and that also is a data that's built by fifteen international partners. So it has a lot of kinds of coordination problems that we would envision here. Not a thousand partners, not a thousand different labs, but at least n jis greater than one.

CHAIRMAN BERG: Well, in the remaining time, since I would like to have us break by not later than four-thirty, let's assume for the moment that we've made the case. We have a budget, and somebody has given us the go-ahead to proceed on this project. It seems to me a very important effort is how one organizes.

MR. HOOD: Can I raise just one point? Have we taken into account developmental costs for technology, Paul?

CHAIRMAN BERG: No.

MR. HOOD: It seems to me -- Are we going to get to that somehow? I've made the point, I think that's a really big part of it. We talked about how much it costs to sequence and how much it costs to map, but we haven't talked about --

MR. COOK-DEEGAN: The only comment that I've heard all day is your suggestion that there should be a certain number of groups from which we could multiply out developmental costs. We assumed that there were six to ten of such of those groups of ten people each. That's the only comment I heard all day.

MR. HOOD: I don't know if we want to discuss it or not, but I think we ought to factor in here --

CHAIRMAN BERG: Well, it's certainly a relevant number to this total. You want to --?

MR. HOOD: Well, I guess my feeling is that, in all of these major areas, the DNA handling and the mapping and the sequencing in this computation, there are these opportunities that could, I think, easily afford ten groups that range in size from two to ten, senior people -- okay? -- that could operate to explore one or two or three or more of these areas. I would think you would want to have these groups widely dispersed, and having interaction with the molecular biological community, and -- I could sit down and give you some numbers instead of, you know -- average numbers for technicians and professionals in each of these groups, but I would think that should be -- I would think you would want to be talking, once you get this thing going, in the order of $50-$75 million a year at a minimum to really push the technology development.

MR. COOK-DEEGAN: Any other comments on technology development?

MR. WATSON: Isn't that pretty large, $75 million?

MR. HOOD: Well, I think other groups have come up with numbers that are similar to that.

MR. WATSON: That's 750 people.

MR. HOOD: Yeah, but I mean, the other thing that we really haven't factored in which has to be included in that is construction costs, all of these other kinds of things that get groups starated and creating the environment in which they can operate.

John, how many people at the MRC Lab of roughly 200 people are technology people?

MR. SULSTON: Well, I wanted to make a comment, yes, exactly of that sort that implies. I think you agree, Lee. I mean, it's the way one operates, isn't it, usually? You pick a problem, and you tackle it, and the technology develops.

I mean, the way I always found it, you know, was in saying what it would cost if we did what we were doing now forty times enlarged. But, of course, it's implicit in that that that's the upper bound, and every scientist operates all the time by trying to do things better. And whether there's actually any identifiable areas where you can kind of take a group and put them separately from doing it and say, develop it, I rather doubt. I think that's what Jim's implication is, that basically the development goes on as you --

MR. HOOD: Oh, I agree with that.

MR. SULSTON: Yes. And that's the way you operate, after all. So I mean, I think it would be a mistake to suggest to OTA that it ought to be structured any other way, because I think that makes a bad science.

MR. HOOD: Oh, I certainly agree with that. What I will say is it does cost money to develop technology.

MR. SULSTON: Oh, yes. That's the thing.

MR. HOOD: Certainly, there isn't much money for developing technology that's generally available.

MR. SULSTON: The thing is, if one is given a budget to sequence part of the human genome or map it or whatever one does, then I would say implicit in getting that grant is that you also include the technique. I'd take that for granted.

MR. COOK-DEEGAN: One way to look at it is this is an add-on cost. Using John's model of how this is actually going to play out, that's incorporated into our ideas of RFLP mapping and physical mapping, and saying that technology is going to spin out of money you're already planning to spin. From what I understood, what you were saying perhaps is some of that should be focused on instrumentation development or something like that that would be an add-on cost to what's already going on.

MR. HOOD: What I've heard us discussing is how much it's going to cost to map or how much it's going to cost to sequence, and I think those estimates probably are production line estimates with current technology. They don't take into account the really expensive technology development.

Now, I agree, that's the way you want to do it, and if we're to do it that way, we should just add on a significant factor to these groups who are going to have sequencing or mapping or whatever it is.

MR. SULSTON: Sorry, I didn't -- No. I say, the point is that, if you develop correctly, then it costs you less to do than it would if you just did it in the current fashion. If that's not the case, then the development was ill advised.

MR. HOOD: Well, actually, our experience is the other way around. It costs just as much. You just do twice as much. I mean, one way or another, you have to have money --

MR. SULSTON: In this project, that comes to the same thing. Right. So you do it twice as fast. You spend your first year developing and your second year doing it twice as fast.

MR. HOOD: So let's go back. So at ten cents a base we're going to have set up groups to do DNA sequencing. It's the general feeling that you could also do the kind of technology development you'd like to see done in DNA sequencing?

MR. COOK-DEEGAN: Actually, I was planning on using the twenty cents a base. What I'm hearing is a very different philosophical view between the two of you about whether -- The argument, as I understand what you're making, John -- you know, you say, yeah, we're going to do the physical map and RFLP map and sequencing, and built into those costs can be an assumption it's going to get cheaper to do that. The development costs are going to saturate how much cheaper it is by being built into our --

MR. SULSON: This is also -- Bear in mind, this is the answer to Paul's problem about the thing being boring. I mean, that's why it's not boring is because you do it better all the time.

MR. HOOD: Let me make the counterpoint, and that is that I think a big part of technology development today is very expensive instrumentation. It is not the simple add-on to DNA sequencing to design and set up a flourescent -- a laser detector or to build machinery that can simultaneously detect fifty channels in four different flourescent modes.

You're talking about engineering, and you're talking about instrumentation development that is way apart from just making the molecular biology a little better. And it's going to be more and more in that direction. I mean, we're getting more and more sophisticated in how we do these things.

So I think it's incorrect to assume that there is no significant add-on to the technology development.

CHAIRMAN BERG: Could we focus on precisely those kinds of technologies that you would see as being -- sounds like the criterion is they're divorced from the mainstream of current molecular biology.

MR. HOOD: I think they shouldn't be divorced, because I think the automation and instrumentation development should be grounded in the molecular biology and what it really needs and the best way to do it. I think that's essential.

CHAIRMAN BERG: Lee, as I understand it, your place is probably the only place in which that's either being done or is possible.

MR. HOOD: Well, there are companies that are doing a lot of these different things.

CHAIRMAN BERG: Yes, but not doing those two in conjunction. Most companies are --

MR. PEARSON: Yes, they are.

CHAIRMAN BERG: Well, okay. So there are exceptions. (laughter) There have been more labs involved in sequencing. There are very few people within that format who are then going to undertake the instrumentation development that's going to make their job easier. They're going to rely on people who got their kicks another way and developed the kinds of instrumentation that helped them. For example, -- Well, I won't get into that.

MR. COOK-DEEGAN: Mark, can you help us out here? I mean, suppose duPont decided to make a major initiative, and a component of it has to be technology development. Do you throw in an extra factor, or do you -- I mean, how do you budget something like that? Help us out.

MR. PEARSON: Well, we do it the same way anybody else does, and Lee's lab, I think, is a paradigm. We started out under the table, because there's a need; and our guys are daunted by the amount of effort required to go ahead and do things in conventional methods and, unlike popular conception, we don't have an infinite resource and a huge number of people. So we look for a cheaper, better, faster way of doing it. I think that simply supports exactly what John is saying, and I know that's how things got started in Lee's lab as well.

MR. SULSTON: It's a sure way to create white elephants, if you have a lab which doesn't have a goal and it's just developing an instrument.

MR. COOK-DEEGAN: Well, I don't think there's any disagreement about that. The disagreement is about how you budget it really. I think, Lee, are you holding the position you need to add some extra money in there?

Even though you expend it at the same places, you need to add some extra money?

MR. HOOD: Well, I do. And again I just make the point, I think to do a lot of this kind of instrumentation -- and Mark will back me up -- you need to bring into these groups engineers and compatuer scientists. I mean, it's an important part of this DNA -- It's the biggest part of the DNA sequencing. Chemistry is a trivial part by comparison in some areas.

So if you're really going to develop these sophisticated new tools, you need to bring in these other people, and that costs money; and it's not a simple add-on just to doing the molecular biology.

MR. PEARSON: I support Lee's view, but I would take exception to the triviality of the chemistry, because in fact there are significant advances, at least we feel, that can be made, will be made -- Well, it's just to expand the context in which this is being stated. What's happened here is we're seeing a whole bunch of fields and levels of expertise applied to a general problem, and they are -- you know, this is hardly a recognizable view of molecular biology as Jim, for instance, knew it twenty years ago.

MR. SULSTON: What you're saying is you must have it in the one lab so you have just as much incentive to improve the chemistry as you do the computing, rather than focusing on one bit of it.

MR. PEARSON: And as we've seen through the technology that's been developed in the last year or two, we've moved the monkey from one back onto another back. It's a different back, but the monkey is still there.

MR. HOOD: I think what we have the opportunity to do here is to create the resources to attract into biology new sets of people that can make enormous contributions. Now we can continue doing biology as we've done in the past, and that would be gradual improvements and it wouldn't be technologically oriented.

I personally think that would be a big mistake. This program, it seems to me, could be a paradigm for bringing together in a real useful and constructive way all these different threads of physics and computer science and engineering and chemistry and biology.

MR. COOK-DEEGAN: The question then becomes, how do you factor that into the budgeting process, and it sounds like one way we could do that would be to, instead of trying to make a flat estimate of costs, do another thing like we're talking about doing with the data base. Essentially, say you add a certain percentage in. What's the percentage?

MR. STEVENSON: Four to eight percent, typically, in industry.

MR. HOOD: I would say that's a factor of three or four.

MR. McKENNEY: If I understand what Lee is trying to say, there's clearly the advancement you get when you're trying to solve a hard problem that John has certainly demonstrated. But I think one of the things is there's a basic research problem that's a long term commitment that you don't get an immediate return on; and that's the 25 percent over and above what you spent to solve the problem.

One of the areas that's obvious to me is it would be nice if we had better restriction enzymes, but to spend three years to solve the structure and engineer them is not going to give a return up front. And nobody is going to pursue those sorts of problems to have the short term return, to improve the technology.

MR. WATSON: Lee, what was the amount of money spent at Cal Tech --?

MR. HOOD: Boy, that number is really hard to come by. I would guess we're now spending a million and a half a year.

MR. COOK-DEEGAN: What percentage of your effort goes to that sort of effort within the lab, in terms of people and thought processes that are --

MR. HOOD: Half -- 60 percent.

MR. CARRANO: I would agree with Lee. I think the estimates we've heard before are much too low. I think the National Laboratories have had a long history of mixing technology development with the biology, and I would say that their efforts in technology development are almost comparable to their efforts in biology. They're almost equally weighted, and at least 50 percent of the costs of the biology and chemistry.

So I think the number of 50 percent is probably more accurate, if not even low.

MR. WATSON: What have you developed --?

MR. CARRANO: I think the National Laboratories have been instrumental in developing the flow systems and have pioneered that technology, for one. They have been instrumental in developing the improvement of lasers, as another.

MR. WATSON: But I still don't have the feeling of whether the National Labs have been efficient or inefficient in doing it.

MR. HOOD: You know, I think the --

MR. WATSON: That's all I'm trying to say. Fifty percent sounds good, but it's not clear to me that per dollar spent it was very well spent.

MR. HOOD: I think the question Jim is also getting at is how many people are there out there that would take advantage of these things? I think that's an important question. But my point is, if the resources aren't there, you'll never find out. I mean, you know, young people get discouraged and do classical kinds of things, because that's where the money is.

MR. COOK-DEEGAN: We're trying to bring this to a close. I mean, we heard the industrial figure is four to eight percent. You're estimating -- multiply that by four. We're oging to end up with something around a quarter of -- 25 percent surcharge or something like that that would be incorporated into the budget, spent the way you were already talking about spending it but having the added mandate of developing technology. Is that how I understand the succession?

MR. WATSON: If you could clone Lee.

MR. GILBERT: I would have thought in investigator originated grants.

MR. PEARSON: There is another way of looking at this. I was recently on a National Science Foundation biological instrumentation panel, and I know that of the requests that were submitted in this general area, and not many of them on DNA sequencing but on instrumentation development for biological science, we probably ended up funding maybe fifteen or twenty percent of the total number of rants that were plugged in.

It's clear that there is a much larger -- That's because of dollar limitations, and it was quite clear that many more could have been supported. So the fact is, there are good ideas going begging out there, because we don't have enough money in our current granting programs to support this level of activity.

CHAIRMAN BERG: Off the top of my head, I find it hard to say that there's three times more need and utility for instrumentation development than there is in the whole computing problem that we were talking about before. I mean, we're talking about 15 percent of the total in terms of all the things that we feel are needed to be able to handle the data and develop new approaches to using it and the hardware and the software and like that. Now you're saying that three times that level would be needed for instrumentation development? Somehow it just -- Somebody said 50 percent.

MR. HOOD: We finally settled on 25.

CHAIRMAN BERG: Oh, you said 25. Sorry. Okay.

MR. McKENNEY: It's not just instrumentation. I mean, it's enzymology, biochemistry, chemistry. Okay? It's all aspects of --

CHAIRMAN BERG: You're the only who's said that. I didn't hear anybody else say that that was included in this thing.

MR. HOOD: I agree with that completely. Sure. It's developing, you know, whatever you want to

MR. COOK-DEEGAN: This is for all four components you were talking about originally?

MR. HOOD: I'm talking about all four components. So I would include -- I mean, maybe you want to sort out the computing development from this, but I'm talking about the four things that I initially discussed.

MR. STEVENSON: Also talking about general improvement of molecular biology resources, not just specifically for this program.

CHAIRMAN BERG: Okay. We're talking about twenty-five percent. So we've got two add-ons of 15 and 25, roughly.

MR. HOOD: Well, and you may want ot factor one of those add-ins, because it's --

CHAIRMAN BERG: It's part of another. Okay.

Let's turn to this last item, which I have assumed meant how is this all going to be managed, in a nutshell. That is, given that there is enthusiasm and deep pockets, that -- and, in fact, we need some protocol for how it's going to be managed (a) by whom, what the mechanisms are. Well, those two, essentially.

There are items there that deal with detailed quality control, training,facilities, the ethical analysis, I don't know if we'll have time for.

MR. COOK-DEEGAN: The person who could handle those isn't here anymore.

CHAIRMAN BERG: Yes. The input from the outside world, or the practicing scientists, into this whole process -- that's the advisory group. I'm going to just start -- Jim, you want to start? I mean, you've been a strong advocate of the NIH as being the logical and probably so, legitimately sole group to take on this kind of a project.

MR. WATSON: I think it's connected to the users of the project and to the people who are most capable of doing it that are traditionally supported by NIH.

CHAIRMAN BERG: Do you think the existing NIH organization can absorb and integrate this kind of a project into its current framework and -- I'm just curious. I mean, the whole management at the NIH and the grants you process -- I almost have the impression --

MR. WATSON: I don't think it would faze them to do that.

MS. KIRSCHSTEIN: Do you want me to say something?

CHAIRMAN BERG: Yes, please.

MS. KIRSCHSTEIN: First of all, I think that NIH and the Department of Energy, EPA, the NSF are all collaborating in whatever is being thought about now in the governmental agencies. We're fully aware of the fact that there are all these articles that say that everybody's buying, and we're not. We're working very well together. Dave Smith, I hope, is still here; but we are working very well.

We attend each other's meetings. We are now setting up and already have set up and have had at least one meeting a group called -- a subcommittee on this project of the biotechnology working group of the Domestic Policy Council.

We have arranged for rotating chairmanships. Dr. Wyngaarten himself is the chairman for the first year, and whoever he or we wishes to designate will be the chairman for the second year. We are exchanging information. We are gathering information about what each of our grantees, contractors, etc. are doing.

Secondly, within NIH Dr. Wyngaarten has a working group. You can see his interest, because even in that working group he himself is chairing it, although when he can't do it he has asked me to do so.

We think that we can work these things out among ourselves, and we think we're doing it well. The thing that will stop it as a project is if there is not a lot of new money. Each of us, DOE and we, each testified in front of Congress about the project this year, and thus far things look encouraging. The House committee has voted $30 million to start a mapping project.

In terms of technology and methodology, I'd like to put in a plug for some things you people don't seem to do. We have made some announcements in the NIH guide to grants and contracts about being interested in receiving research grant applications for just the kinds of things that Lee was talking about. They came out last June. I talked to most people around here. They haven't even seen them.

We sent them out to people who don't bother to read them when we send them out to them. I'm not trying to be crass about this, but we would like your input, and people don't seem to be realizing that we want to do this. I don't know how to get the word across.

That doesn't say that we're immediately going to stop doing everything else. I think the thing that is concerning NIH is that there is a great deal of other good science that needs to be done; but we want to do some of each, depending on how much money will depend how much of it.

MR. COOK-DEEGAN: Setting aside the how and all that stuff, I guess one of the questions that it would be most helpful to answer in this context is, assuming -- We're kind of assuming the whole exercise of today has been assuming an incremental jump in the amount of money flowing through existing streams or creating new streams or whatever. There are administrative costs associated with that. Is there any rule of thumb for estimating what those -- Like, presume that there are going to have to be a few more grant administrators to be able to funnel the money around and a few study sessions and things like that.

MS. KIRSCHSTEIN: Aside what Dan Macies was talking about, I do not think there will have to be very many more grant administrators to handle the projects. If one were talking about what Dr. Gilbert was talking about, which is to do some things like doing it under a contract for quality control, it is the view of many of us that people in the government who are the project officers for that have to watch those kinds of efforts much more carefully, and that takes a much more intensive amount of manpower.

I think all of us can see -- I can tell you that I've had two people working full time on the GenBank contract for a very long time. So -- but for grants, for investigator initiated ideas, even the kinds of things that we ask for in these announcements, I don't think it will take a lot more manpower.

MR. COOK-DEEGAN: So we're talking about less than a couple of percent or something?

MS. KIRSCHSTEIN: Yes.

MR. HOOD: But doesn't that depend on the -- Suppose that you were given $150 million a year increase for mapping and sequencing the human genome. Would you think -- you would just distribute that to the institutes at NIH and then handle it in a present administrative --

MS. KIRSCHSTEIN: Probably.

MR. HOOD: And you wouldn't think it would be advisable to have a separate -- separate study sections and/or administrative structure.

MS. KIRSCHSTEIN: Even if there were special study sections and we have a special study section for this, we might need one or two extra people. I don't really think that that's a big concern. Would the staff agree with me?

CHAIRMAN BERG: Ruth, a common charge or concern that's expressed -- been expressed on numerous occasions, and that is if the commitment were made to do this and the NIH were given that mandate, what's the guaranty that that mandate will remain in effect over the lifetime of the project in the sense that the incremental funds would continue in a stream rather than at some point under some economic crunch of saying we don't have anymore funds; if you want to run this project, you've got to take it out of existing NIH funds? I mean, that's a critique which has been made.

MS. KIRSCHSTEIN: I don't think that's any different, no matter which government agency does it.

CHAIRMAN BERG: I think that's a charge -- I mean, that's a criticism about the whole project, frankly.

MS. KIRSCHSTEIN: The current economic situation is that --

CHAIRMAN BERG: The NIH, I think -- Having it in the NIH makes the whole NIH program more vulnerable. If it's in the DOE, then the DOE worries about it. Right? I mean, the people in the government labs, but if it's in the NIH and then suddenly there's -- you're $150 million or $200 million a year committed, and suddenly that flow stops. You have to get it from somewhere, because these labs are in operation. Where do you get the money from?

MS. KIRSCHSTEIN: Well, that is something that has concerned all of us and, I think, concerns Dr. Wyngaarten as well. It is NIH's general philosophy that much of the research we are doing -- in fact, all of the research we are doing is extraordinarily important, and that there is a great deal of research even in genetics which -- excellent research which we cannot fund today.

I don't think any of us want this project to take over in a way that would stop the rest of it. I think, without indicating it, you can see what is happening in some other areas these days; but if there were no funds, or even if there aren't new funds, at least knowing how the scientific community feels about this project, NIH would have to do some work.

Actually, as you know, work related to what I like -- prefer to call characterizing complex genomes, and that means all of them, right now NIH is supporting $300 million in that regard. It has been for quite a while. It's not specifically, but a lot of it is mapping, and a lot of it is sequencing of genes of interest at the moment. But to take money from other important projects if there are no incremental funds would give all of us very serious problems.

CHAIRMAN BERG: Yes. And I suppose the concern can be expressed initially, because that creates certain reservations about whether you get into it. I think the critique I'm referring to is that three years down the line or four years down the line when you have a major commitment and already widely in progress, and then suddenly the brakes get put on. Then the decision making process becomes really strained.

MS. KIRSCHSTEIN: We have great concerns, but I think so would DOE, frankly, unless I'm badly mistaken.

MR. CARRANO: I think one solution to that is that is that you won't disperse it among several agencies and have an interagency group as the board, the governing board of this effort, and to not create major new institutes to do this but to funnel appropriate monies into existing institutes or create some small new ones, but at lease disperse the money so that if monies do stop years down the line you're not faced with a situation as dramatic as you suggest.

CHAIRMAN BERG: Well, okay. Jim, did you want to --?

MR. WATSON: Well, I was -- had talked to Jim Wyngaarten about it, and I talked to several of the Congressional staff, as well as to several Representatives and Senators. We made the point that we thought it should be an add-on, and we also made the point we thought it should be a line item possibly in the Office of the Director of NIH so that it was quite clear that it was a separate component which wasn't to be coming out of TMS's funds.

MS. KIRSCHSTEIN: What happened was that there was an add-on. It was -- Dr. Wyngaarten was asked -- He originally put it in the Office of the Director, and when he was finally asked, he said he suggested that it be put in my institute. However, I think he's fully aware of the fact that whatever is going to be done, it's going to be in consultation with him. That's the way we're working on this project at the moment.

Everything I say represents basically what he would like to say, I think. I've had enough conversations with him to know that.

MR. WATSON: But we wanted to identify it as a separate thing rather than just part of the --

CHAIRMAN BERG: I mean, safest is you create a trust fund. You put three billion bucks in a trust fund, and you say that's -- and it can't be used for anything else. I don't know how you can manage that, but that seems to me one way you would be sure that there is a steady stream.

MR. GILBERT: I think that this is a goal oriented project and needs to be sufficiently clearly identified so that you can measure it against its goals.

MS. KIRSCHSTEIN: We'll be tracking what we are doing very carefully.

MR. GILBERT: Fundamentally, when it's a granting structure, it ceases to become a goal oriented structure because that's not the nature of the grant.

CHAIRMAN BERG: I mean, the space program is a case in point. I mean, it's clear that that had a mission. I think it progressed along a track that was planned. It was successful, and then at some point the thing began to have to be contracted, and some very severe decisions had to be made and lots of people were very unhappy and are unhappy about some of the decisions that were made. There's a perfect case in point where it was a national commitment, and then somewhere along the line the baton got passed to somebody who didn't share that same commitment, and then you have a slightly different way of thinking.

MR. GILBERT: You need commitment on the part of the agency that takes up the project to do the project. The project needs that well defined goal, some way of measuring that it's getting there.

MR. COOK-DEEGAN: We're now getting into the issues that are the next stage of what we're thinking about in fact. I don't want to cut it off too short, but on the space program analogy, for example, it's -- One thing that's nice about this is that it actually does ramify into existing, ongoing research. So I think it's unlikely to have man on the moon, you don't know what to do. You're going to have a bunch of technologies and information that's widely useful.

Anyway, there are two things that come to mind that could be considered administrative costs that we haven't talked about that I think might be significant, nonetheless. I think we've heard that the add-on costs for the administration in most of these categories is not that big of a deal, but what about personnel and training and quality control. Are there any things that we need to add in to an overall budget in terms of quality control beyond what we've already plugged into the data bases and the instrumentation and stuff like that. Are there any inner group quality control issues and anything, special training programs, things like that?

MS. KIRSCHSTEIN: Let's talk about research training for a moment, because that I do think is important. I think it should be of significance for you to know that over the last really close to twenty years nobody has added any funds to NIH for research training. There has been somewhat of a less than great interest in research training.

We are continuing to do everything we can. So if one really thinks that research training is needed in these areas, then that should be an emphasis for you, because otherwise people are not particularly interested, including the scientific community which has gone to bat over a number of years for other parts of the budget but not for research training.

MR. HOOD: You know, one of the problems we run into, Ruth, is there's been a -- at least as I perceive it -- a bias at NIH for training MDs as opposed to PhDs; and this is a porject where PhDs really are the essence.

MS. KIRSCHSTEIN: The vast number we train are PhDs, and indeed Dr. Wyngaarten has been very concerned because we were training many more PhDs than MDs these days; and he would like to see that somewhat reversed.

MR. CAHILL: I believe the problem is the other way around, as Ruth says.

MS. KIRSCHSTEIN: Yes, exactly.

MR. CAHILL: The MD coming through the training program is the rare bird these days, except those 700 in the MD-PhD programs, and that's a drop in the bucket for the national effort.

CHAIRMAN BERG: Isn't that what Lee is arguing?

MR. CAHILL: That's such a small segment in the total training program on the MD-PhD side.

MR. GILBERT: How much pressure is the training program under now?

MS. KIRSCHSTEIN: A fair amount.

MR. GILBERT: And a real effort to cut it out?

MS. KIRSCHSTEIN: No, but no effort to emphasize it.

CHAIRMAN BERG: Well, I can't imagine that there's anybody here who would be opposed to including training as almost a line item in any kind of recommendation.

MR. COOK-DEEGAN: What I'm wondering is, I've heard several times how many people are there that can do X,Y or Z thing. Seems to me, if that's a question that is raised in connection with the things we're talking about doing on this project only, that's what I would like to restrict discussion to; then it sounds like that is an automatic feature that you have to build into your program.

Now how do you do that, and how much do you throw in that direction?

MR. PEARSON: I think that's absolutely dependent on what we intend to do with the sequence as it's done. From my perspective, it seems that the medical applications of direct sequencing are going to be very large downstream, and we're going to need people in clinical situations who are able to do sequencing.

Now the question is, is this a training vehicle that will be useful for creating that kind of skill?

CHAIRMAN BERG: You're going to have machines.

MR. PEARSON: But as we've just discussed, the machines are one part of it, and perhaps a small part of it.

MR. HOOD: You know, one kind of person that would interpret these, though, Paul, would be people who really know molecular biology well and who know computing well; and that intersection has almost a zero population right now. That's, I think, an area that is going to be terribly important for this area, and people who have, on the one hand, really good organic chemistry background blended together with molecular biology are terrifically useful. So --

CHAIRMAN BERG: Are there a couple of shortage areas like that that we could list, a half dozen or something like that, and specify how many people we need to train?

MR. HOOD: I could certainly give you areas where there's a real shortage. The total numbers are somewhat arbitrary, but we could come up with some numbers.

MR. CAHILL: I think Mark has touched on the very limited supply. Coming up now are the MDs who are conversant with molecular biology.

CHAIRMAN BERG: But not just to do sequencing. I mean, it's to understand basic molecular biology enough to think in terms of molecules when it's talking about disease.

MR. COOK-DEEGAN: I would like to leave that. We're going to get some meaningful feedback to go both to Christian and to us. If we could get some sort of agreement within a week, how many could work -- Lee, would you take that on and list a few things and how many people you think -- shortage areas. That would be very helpful. Quality control is in there.

CHAIRMAN BERG: I assumed what you meant by quality control is not determining whether the sequence is correct but whether the people who are getting funding are appropriately chosen and that there's some oversight.

MR. COOK-DEEGAN: No, I didn't mean that. No, because I kind of think of that as being built into the administrative structure and advisory council.

CHAIRMAN BERG: But again, if I remember correctly at Cold Stream Harbor and that discussion, Jim, that was a major concern about where that authority would be left. That is, who was going to be reviewing the proposals and what kind of quality control mechanisms would they use? Would they be the traditional peer review process, or would they be contractual with some small group deciding what they want to do? Those are the kinds of --

MR. WATSON: I think another real issue is if you're building a large accelerator, one scientist is in charge of it, someone who stands for the program. Are we going to have a large program run by a committee, or is there going to be one person who is in charge of it? I instinctively believe one person should be in charge of it who understands the scientific issues and who is not chosen purely to be an administrator.

CHAIRMAN BERG: But the NIH model says there isn't.

MR. WATSON: I think --

MR. COOK-DEEGAN: Again I would like to -- I think those are very important discussions and we are committing to finding some more out about administrative structure and things like that, but that's not a budget question because you're only talking -- That's a structure question. I will say that -- To clarify what I said, I think that I am presuming a two level, peer review system that DOE has said they're going to use, that's very analagous to what NIH already uses. And we've heard from Ruth that one to two percent is enough to cover the costs of handling the grants administration part of it.

So I'm not saying it wouldn't be peer reviewed but I'm saying we've already plugged those costs in, for convening the groups and things like that. What I'm wondering about is if there are certain features of this where an NBS style -- As you're doing the mapping and there's inter-group differences in techniques and things like that, is there something, some portion of the budget that needs to be devoted to making sure that the methods are mutually compatible and that the mapping information or the clones are maintained in a way that's --

MR. GILBERT: I think you do have to sort of face up to an underlying problem that goes back to what is this really about. I don't think that this can be done -- that you will get the mapping and sequencing by solely using the grant structure; because if you try to use the grant structure with the mapping, what you're inviting is investigator originated different ways of mapping. And the emphasis is, of coure, from the scientist, I'm doing it differently than everybody else is, and I'm really proving that yours is better, but certainly not just going along and doing a little more of the same thing. Whereas, the problem in the mapping is to have a map constructred with pieces that are all the same that you can then use.

The same thing is true with sequencing. If you put out grants for the sequencing, you will get lots of development of instrumentation, new ways of sequencing, test runs on this, test runs on that; but no one is going to take a grant to do the major sequencing.

So in both the mapping and the sequencing, once you've developed the method, you then need another structure which you can then have the thing done.

MR. WATSON: Is that quality control?

MR. GILBERT: It means special administration. It is separate from the other. It's a different quality control.

MR. WATSON: Could I ask John, do you think that you would need one person in charge? I mean, from your own experience on a much smaller project -- I mean, if you didn't exist, do you think your mapping would be finished? I mean, doesn't one person really have to finish up that last ten percent or live and die for the thing?

MR. SULSTON: Yes, I expect so. I mean, if -- (laughter) That's really what I wanted to talk about with YACs, you see; because that makes -- you can then think about the thing globally. Alternatively -- Coming back to mapping now, not sequencing. Okay? -- some kind of a global map.

Without the YACs or maybe in parallel with them, you can have your political subdivision into chromosomes by the sorting procedure, and then -- In fact, the quality controllers can say bad. You're not actually having to compare raw data across labs. You are simply asking each lab to produce a map for certain defined standards. The same thing is true, I think, of sequencing.

Sequencing can be split up, given that you have the clones. You can split it up. It's then a final output quality control. But what I think would be very difficult to set up, and I don't see how you could ever do it, would be to have a number of labs generating raw data on the same shutdown program, either for mapping or sequencing, and fitting it all together. That, seems to me, to be terribly difficult.

MR. GILBERT: Without that, in fact -- Without certain aspects of that, you never have the quality control. If you use a grant procedure, somebody sequences a chromosome. What is the quality control?

MR. SULSTON: Okay. So you need two people to sequence each -- Yes. Basically, yes.

MR. GILBERT: You need some way of looking at it.

MR. SULSTON: Maybe I'll withdraw somewhat from that, yes. I think in sequencing, perhaps it is reasonable to fit them together. The mapping I see as very difficult.

CHAIRMAN BERG: Did I understand? I'm puzzled, because what John is responding and what I thought you said are two different things. You were talking about the whole project director or head, something like a guy who headed up a whole mass of programs. Is that what you're talking about that you need, or are you talking about the level that John is talking about?

MR. WATSON: I think you're going -- to keep it on course, it would be useful to have one scientist in charge of the program.

MR. SULSTON: For the whole thing?

CHAIRMAN BERG: See, that's why -- Your response was to --

MR. SULSTON: I was responding to something rather different. Yes, I was responding on sort of a practical level, really, rather than a philosophical level. Well, I mean, I guess the answer is yes. I mean, if you do have somebody with a dream, then of course, it drives itself. Yes.

MR. WATSON: You're going to get bad years, and you know, the program is going to be under attack, bad economic attack; and there's got to be someone there who sort of sees all the components and fights for it.

MR. SULSTON: Jim, the reason I responded the way I did is that I can see for the way things are and for what you can conceivably generate shortly, to have the sort of lab level -- I mean that's why I was saying break it down, if possible, and so that somebody has to do all the sequence a chromosome or something. Or in the case of Wally who's successful and he has the dream to sequence the whole lot. Right? I mean, that's his dream.

I think it's got to be within that -- Do you see what I mean? I mean, it seems to me rather appalling and I think not really will create so many enemies wanting to really create a biological mess. He's administrating this huge empire. Do you really want to see that?

MR. WATSON: I think someone has got to do it.

MR. SULSTON: You would like to do it.

MR. WATSON: It's a question whether it's an active scientist or a retired scientist.

MR. SULSTON: What I see -- The way I see it most likely going forward is at some point, of course, it becomes easy enough so that the whole human genome is sequenced in one lab in five years. I don't know at what point we'll be at that level of technology, but before that you can have individual dreams -- do you see what I mean? -- which -- Surely, that's the way to do it.

MR. WATSON: Well, I couldn't think of a job I'd like less. (laughter) I'm very relieved. (laughter) But I just have a feeling that someone has to keep track of the whole thing over the long period of time with a certain degree of intensity, or it can sort of slip down and be run as a sort B effort instead of an A effort.

MR. STEVENSON: You know, my understanding of the kind of job that you want to do, with all due deference to Dr. Watson, I think I would consult Lee Iacocca for how to get the job done rather than us; because it seems to me to be one of carefully sorting out which elements of work are appropriate to be done by what mechanism and who is going to coordinate the overall job.

CHAIRMAN BERG: Why do you think Jim couldn't have run Chrysler? (laughter)

MR. STEVENSON: I don't think he would be motivated to do so. I don't know that he might not do a good job if he put his mind to it, but I don't think he would be motivated. I think that's one of the problems here. I don't think there's a grantee who would apply to general medical sciences, in many cases, who's going to apply for the kind of work that Dr. Gilbert says is going to be required. I think --

CHAIRMAN BERG: Well, I don't see that what Jim has in mind as being any different than the man who is NASA Administrator. I mean, they're very much -- I mean, when you started the whole national space program, many of the kind of elements that must have been required for that and planning for it and organizing for it are not different -- They're two different things, but I don't think it's qualitatively very different.

MR. WATSON: It could be an oversight committee with -- you know, at least a chairman. The thought that it would rotate each year with DOE being chairman one year and NIH, it just seems to me that --

MR. HOOD: It would seem to me what would be a good idea would be to have a scientific advisory body.

MS. KIRSCHSTEIN: Well, that's almost inherent in anything we do.

MR. HOOD: Right. Perhaps the head of that could be this kind of director who would have close communication with the agency.

MS. KIRSCHSTEIN: That's very obvious. There's no question about that. In fact, on one of the NIH committees that already -- the internal NIH committees Dr. Wyngaarten has added -- He's already working on that.

MR. WATSON: That sounds good, but he has no qualifications for the job at all.

MS. KIRSCHSTEIN: Well, he thinks pretty well.

CHAIRMAN BERG: We've got five minutes, and I want to ask Bob Cook-Deegan whether he has any unanswered questions that we might hit. If not, we'll --

MR. COOK-DEEGAN: I'll use the last minutes to be pragmatic, and I will say that the questions you are raising at this very last session are precisely the ones that are the next set of questions that have to be addressed legislatively. They're being thought about. It's not a budget question. We will -- We're thinking about how to address that question already.

There may be another workshop on that, or there may be something else that would address advisory structure and organizational aspects to this.

Second, if we could get the materials that people have said they either have in hand or whatever sent to us -- Lee has promised to send something. Mark, I need your Xeroxed notes, and Christian, I need your stuff to be Xeroxed for us. Anybody else in the audience or here, that would also be helpful.

Aside from that, I think -- Well, we'll iterate this process by sending out the transcript to all the panelists and you can send comments following that, if you wish. Then Christian and I will be independently writing something out.

CHAIRMAN BERG: I want to say that, when Bob asked me to chair this meeting and also when I saw the agenda, I thought there wasn't a snowball's chance in hell that we would actually come up with as much as I think we have in this discussion.

So I want to thank each and everybody here for their participation and thoughtfulness, and wish you good luck. It strikes me that, from a year ago at Cold Spring Harbor to where I'm listening to things now in the last few weeks, it's been an enormous change in (a) the direction or the extent to which the thinking has progressed, and also the lessening of the din of the disbelievers or the people who were strongly opposed.

I haven't heard as much from them as I did at that meeting, and I don't know whether they are still out there; but it seems to me that the thing is going a lot further than i would have thought it had gone from a year ago.

So, thanks again.

(WHEREUPON, the workshop was adjourned at 4:30 p.m.)

- - -

CERTIFICATE

This is to certify that the foregoing transcript

In the matter of: WORKSHOP: "COSTS OF HUMAN GENOME PROJECTS"

Before: OFFICE OF TECHNOLOGY ASSESSMENT

Date: AUGUST 7, 1987

Place: WASHINGTON, D.C.

represents the full and complete proceedings of the aforementioned matter, as reported and reduced to typewriting.

Miles Anderson
MILES ANDERSON